Unit 1
Descriptive
Stastics
Course Pack

Introduction to Statistics

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.
- One in every 8 South Africans is HIV positive.
- By the year 2020, there will be 15 people aged 65 and over for every new baby born.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. For example, consider the following scenario and the interpretation based upon the presented statistics. You will find that the numbers may be right, but the interpretation may be wrong.

Example 1: Try to identify a major flaw with the following interpretation before we describe it.

> *A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.*

_____
_____
_____

As a whole, the above example shows that statistics are not only facts and figures; they are something more than that. In the broadest sense:

Statistics refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.

Students study statistics for several reasons:

1. Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.

2. You may be called on to conduct research in your field, since statistical procedures are basic to research. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. You must also be able to communicate the results of the study in your own words.

Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases.

Type of Statistics
Broadly speaking, applied statistics can be divided into two areas: descriptive statistics and inferential statistics.

I. Descriptive Statistics

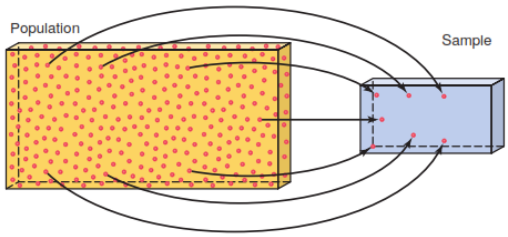
Suppose we want to have an idea about how well do Bayview students do in Stat unit test in the past 5 years. In statistical terminology, the whole set of numbers that represents the scores of students is called a **data set**, the name of each student is called an **element**, and the score of each student is called **an observation**. A data set in its original form is usually very large. Consequently, such a data set is not very helpful in drawing conclusions or making decisions. It is easier to draw conclusions from summary tables and diagrams than from the original version of a data set. So, we reduce data to a manageable size by constructing tables, drawing graphs, or calculating summary measures such as averages. The portion of statistics that helps us do this type of statistical analysis is called **descriptive statistics**.

Definition:

Descriptive Statistics: consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.

II. Inferential Statistics

In statistics, the collection of all elements of interest is called a **population**. The selection of a few elements from this population is called a **sample.** The practice of selecting this subset is called **sampling.**



One example of the difference between a sample and the population is seen when governments hold an election. Numerous researchers ask people who they will vote for as the election day approaches, but they clearly cannot ask every voter. They use a sample. On election day, the government collects the official votes from the entire voting public (or at least those who choose to vote). They are collecting data from the entire population.

Another example of sampling could be seen in a study of the impact of tainted water on the people of a particular region of the world. It would be impossible to find and speak with every person who lives in the region, let alone get their consent to run medical tests on them to collect data for the entire population you want to study. However, you could travel throughout the region to find a sample of people who vary in age, size and sex in order to test the effects on a representative cross-section of the population.

How large a sample do you need?

There are some rules for how large a sample should be for specific statistical tests, but it is hard to generalize. As long as the sample represents the relevant features of the full population, it is a good sample. The larger the sample, the better it represents the full population. There is no set rule regarding how big your sample should be. However, you will notice that on reading a published statistical study, it will probably state how much data was collected and the method of collection. It

seems a little unfair, but it is unlikely that you will be criticized for having too much data. However, you may be criticized for not having enough data.

Definition:

Inferential Statistics: consists of methods that use sample results to help make decisions or predictions about a **population**.

Sampling methods

One characteristic that every good sample has is that it is **random**. This means that each potential data point has the same probability of being chosen. There are several ways to select your sample. Some are shown in the table below.

Types of random sampling [1]		
Type	**Description**	**Examples**
Simple	Achieving randomness by a simple, completely random process.	You choose items out of a hat or you use a random number generator to pick items from a list.
Convenience	Choosing a sample based on how easy it is to find the data.	You ask the first twenty people who walk past you to answer your survey.
Systematic	If data is listed, selecting a random starting point and then choosing the rest of the sample at a consistent interval in the list.	You roll a die and get a 6, so you start with the 6th item and then choose every 10th item in the list after that.
Quota	Choosing a sample that is only comprised of members of the population that fit certain characteristics.	You want a sample of 50, but they must all be ladies between the ages of 16 and 25.
Stratified	Choosing a sample in a way that the proportion of certain characteristics matches the proportion of those characteristics in the population.	The population is 45% male and 55% female, so you make a sample of 100 people that has 45 men and 55 women in it.

Question 1
Nathanael wants to research the amount of sleep that students in years 9–12 get each night. To collect his data, he goes to several different tables at lunch and asks whoever is sitting there. Identify which kind of sampling Nathanael is using.

a) Convenience
b) Systematic
c) Quota
d) Stratified

[1] Watch the following vide . https://www.youtube.com/watch?v=be9e-Q-jC-0 . This video describes five common methods of sampling in data collection. Each has a helpful diagrammatic representation.

Question 2
Nathanael decides to correct his sampling strategy by making sure each class is represented in the sample according to the percentage of the school they comprise. State the type of sampling Nathanael is using now.

a) Stratified
b) Simple
c) Convenience
d) Systematic

Practice
Attendance and Grades
Read the following on attendance and grades, and answer the questions.

A study conducted at Manatee Community College revealed that students who attended class 95 to 100% of the time usually received an A in the class. Students who attended class 80 to 90% of the time usually received a B or C in the class. Students who attended class less than 80% of the time usually received a D or an F or eventually withdrew from the class. Based on this information, attendance and grades are related. The more you attend class, the more likely you will receive a higher grade. If you improve your attendance, your grades will probably improve. Many factors affect your grade in a course. One factor that you have considerable control over is attendance. You can increase your opportunities for learning by attending class more often.

1. What are the variables under study?

2. What are the data in the study?

3. Are descriptive, inferential, or both types of statistics used?

4. What is the population under study?

5. Was a sample collected? If so, from where?

6. From the information given, comment on the relationship between the variables.

7. Explain whether each of the following constitutes a population or a sample.
   a) Number of personal fouls committed by all NBA players during the 2008–2009 season
   b) Yield of potatoes per acre for 10 pieces of land
   c) Weekly salaries of all employees of a company
   d) Cattle owned by 100 farmers in Ottawa
   e) Number of laptops sold during the past week at all computer stores in Toronto

Answer
1. The variables are grades and attendance.
2. The data consist of specific grades and attendance numbers.
3. These are descriptive statistics.
4. The population under study is students at Manatee Community College (MCC).
5. While not specified, we probably have data from a sample of MCC students.
6. Based on the data, it appears that, in general, the better your attendance the higher your grade
7. a. population b. sample c. population d. sample e. population

Variables and Types of Data

Variables can be classified as **qualitative or quantitative**. Qualitative variables are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (male or female), then the variable *gender* is qualitative. Other examples of qualitative variables are religious preference and geographic locations.

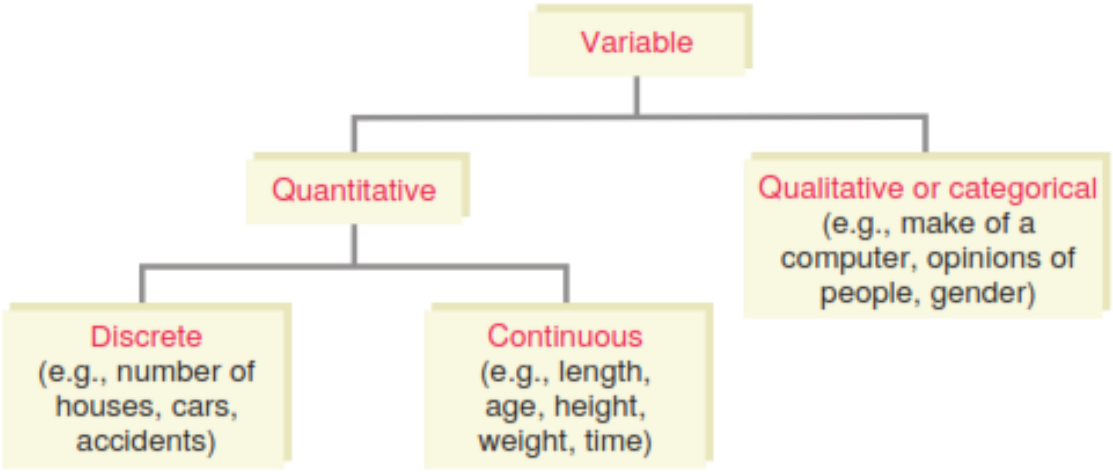
Quantitative variables are numerical and can be ordered or ranked. For example, the variable *age* is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures.

Quantitative variables can be further classified into two groups: discrete and continuous.

Discrete variables can be assigned values such as 0, 1, 2, 3 and are said to be countable. Examples of discrete variables are the number of children in a family and the number of students in a classroom.

Continuous variables, by comparison, can assume an infinite number of values in an interval between any two specific values. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures.

The classification of variables can be summarized as follows:



Practice

1. Indicate which of the following variables are quantitative and which are qualitative.
   a) Number of typographical errors in newspapers
   b) Monthly TV cable bills
   c) March break locations favored by high school students
   d) Number of cars owned by families
   e) Lottery revenues of Canada

2. Classify the quantitative variables in above example as discrete or continuous.

Answer

1. a. quantitative b. quantitative c. qualitative d. quantitative e. quantitative
2. a. discrete b. continuous d. discrete e. continuous

Reliability
The first question is primarily concerned with the reliability of your data. **Reliable data is free of errors and bias**. While no research produces data that is perfect, reliable data comes from individuals who do their best to perform their research carefully, responsibly and without attempting to manipulate the results to support a particular conclusion. Data can be unreliable when individuals are careless when collecting or recording information.

Bias is another type of unreliability.

A bias is one in which the data has been unfairly influenced by the collection process and is not truly representative of the whole population

Consider an online advertisement that says, 'What do you think of Politician X's plan to combat terrorism? Click here to complete a survey.' In this example, bias can occur accidentally because people are more likely to participate if they have a very strong opinion about that politician or about the issue. This means that a large portion of a population probably goes unstudied because they simply do not care enough to take the survey. Bias could also be more blatant if this survey were only advertised on a website run by supporters of Politician X or if the question were worded like this: 'What do you think of trusted Politician X's clever plan to combat terrorism?'

Outliers
The second question is concerned with the concept of outliers. Outliers are data values that are very different to the rest of the data. They can occur for a number of reasons. An outlier might simply be a naturally occurring extraordinary value, such as a student who did not study scoring 20% on an exam. On the other hand, it may be a value that is the result of abnormal circumstances, such as a runner who takes over a minute to finish the 100 m dash because they twist an ankle and limp to the finish line. You will learn how to identify outliers and show them on a box-and-whisker plot later in this unit.

Organizing Quantitative Data

When data are collected, the information obtained from each member of a population or sample is recorded in the sequence in which it becomes available. This sequence of data recording is random and unranked. Such data, before they are grouped or ranked, are called raw data.

Definitions

Raw Data: Data recorded in the sequence in which they are collected and before they are processed or ranked are called raw data.
Frequency Distribution: A table that lists all the categories or classes and the number of values that belong to each of these categories or classes.

Example 2: Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. In this case, these ages are listed in *Forbes Magazine* and are listed next.

49	57	38	73	81
74	59	76	65	69
54	56	69	68	78
65	85	49	69	61
48	81	68	37	43
78	82	43	64	67
52	56	81	77	79
85	40	85	59	80
60	71	57	61	69
61	83	90	87	74

Since little information can be obtained from looking at raw data, the researcher organizes the data into what is called a ***frequency distribution***.
Complete the table:

Class boundaries	**Tally**	**Frequency**
37≤a<45		
45≤a<53		
53≤a<61		
61≤a<69		
69≤a<77		
77≤a<85		
85≤a<93		
	Total	

In this distribution, the values 37 and 45 of the first class are called *class boundaries*. Lower boundary = 37 and upper boundary =45. Note that in the above table, when we write classes using class boundaries, we write to less than to ensure that each value belongs to one and only one class. As we can see, the upper boundary of the preceding class and the lower boundary of the succeeding class are the same.

Constructing Frequency Distribution Tables
When constructing a frequency distribution table, we need to make the following three major decisions.

1. Number of Classes
Usually the number of classes for a frequency distribution table varies from 5 to 30, depending mainly on the number of observations in the data set. It is preferable to have more classes as the size of a data set increases. The decision about the number of classes is arbitrarily made by the data organizer. As a rule of thumb, statisticians use the following formulae.

One rule to help decide on the number of classes is Sturge's formula: $C = 1 + \lceil 3.3\log(n) \rceil$ Where $\lceil x \rceil$, ceiling function is the smallest integer greater than or equal to *x*, **C** is the number of classes and **n** is the number of observations in the data set. Another rule is $C = \sqrt{n}$.
In Example 2 n=50, hence

$$\begin{aligned} C &= 1 + \lceil 3.3\log(50) \rceil \\ &= 1 + \lceil 5.6 \rceil \\ &= 1 + 6 = 7 \end{aligned}$$

2. Class Width
Although it is not uncommon to have classes of different sizes, most of the time it is preferable to have the same width for all classes. To determine the class width when all classes are the same size, first find the difference between the largest and the smallest values in the data. Then, the approximate width of a class is obtained by dividing this difference by the number of desired classes.

$$\textbf{Approximate class width: } W = \left\lceil \frac{\textbf{Largest value} - \textbf{Smallest value}}{\textbf{Number of classes}} \right\rceil$$

In Example 2, we got C=7, therefore

$$W = \left\lceil \frac{90 - 37}{7} \right\rceil = 8$$

3. Lower Limit of the First Class or the Starting Point
Any convenient number that is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

The **class midpoint** X_i is obtained by adding the lower and upper boundaries and dividing by 2, or adding the lower and upper limits and dividing by 2:

$$\boxed{X_i = \frac{\textbf{lower boundary} + \textbf{upper boundary}}{2}}$$

Relative frequency of a class = $\frac{f}{n}$

Percentage of values in each class = $\frac{f}{n} \times 100\%$

where f = **frequency of the class** and n =**total number of values**
Constructing a Grouped Frequency Distribution-Summary

Step 1 Determine the classes.

- Find the highest and lowest values.
- Find the range (***R* = highest value - lowest value**)
- Select the number of classes desired.
- Find the width by dividing the range by the number of classes and rounding up.
- Select the first lower class boundary (usually the lowest value or any convenient number less than the lowest value)
- Find the first upper class boundary

Step 2 Tally the data.
Step 3 Find the numerical frequencies from the tallies

Sometimes it is necessary to use a *cumulative frequency distribution*.

Cumulative Frequency Distribution: A cumulative frequency distribution gives the total number of values that fall below the upper boundary of each class.

The values are found by adding the frequencies of the classes less than or equal to the upper class boundary of a specific class. This gives an ascending cumulative frequency. Cumulative frequencies are used to show how many data values are accumulated up to and including a specific class. On a cumulative frequency curve, points plotted will be of the ordered pair:
(upper class boundary, cumulative frequency). This point represents the number of data less than the endpoint of the interval.

Example 3: The following data give the total number of iPads sold by a mail order company on each of 30 days. Construct a frequency distribution table.

8	25	11	15	29	22	10	5	17	21
22	13	26	16	18	12	9	26	20	16
23	14	19	23	20	16	27	16	21	14

Complete the table:

Class boundaries	**Mid-height(x_i)**	**Tally**	**Frequency**	**Cumulative frequency**
	Totals			

Graphing Grouped Data

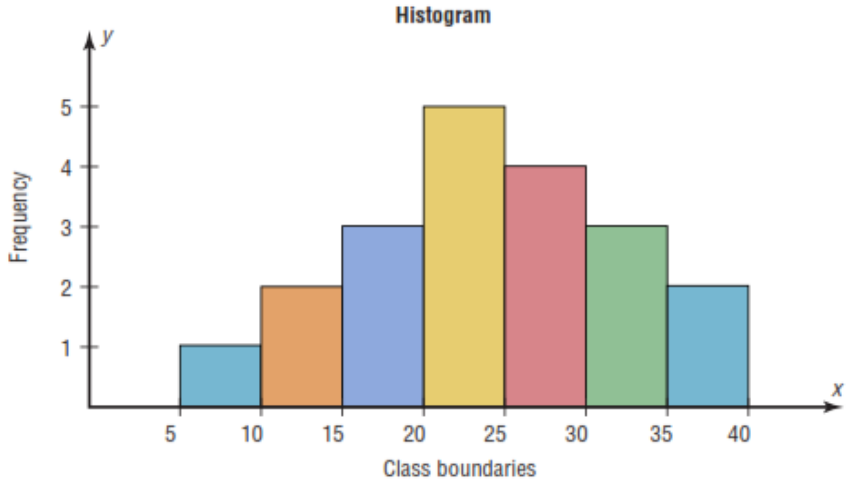
Grouped (quantitative) data can be displayed in a **histogram**.

Histogram: A histogram is a graph in which classes are marked on the horizontal axis and the frequencies, relative frequencies, or **frequency density** are marked on the vertical axis. The frequencies, relative frequencies, or frequency density are represented by the heights of the bars. In a histogram, the bars are drawn **adjacent to each other**.

- It is useful to work out the **frequency density** for each class, which is defined as:

Area = frequency = Frequency density × class width

Frequency density $= \frac{f}{w}$, where f = frequency and w = class width



Example 4: 100 students take a mathematics test that has a maximum possible score of 40. The results are shown in the table. Show the results on a **frequency density histogram** using GDC.

Solution: Calculate the frequency densities.

Mark	Mid-Mark x_i	Class width (w)	# students (f)	Frequency density $\frac{f}{w}$
$1 \le m < 6$			3	
$6 \le m < 11$			10	
$11 \le m < 16$			22	
$16 \le m < 21$			28	
$21 \le m < 26$			20	
$26 \le m < 31$			10	
$31 \le m < 36$			2	
$36 \le m < 41$			5	

TECHNOLOGY INSTRUCTION **Organizing Data**

GDC INSTRUCTIONS FOR HISTOGRAMS:

For RAW DATA

To enter the original data, press the STAT key and choose 1 for EDIT from the screen menus. If necessary, press 4 followed by the keys L1 (2^{nd} 1), L2 (2^{nd} 2) etc ENTER to clear any previous lists.

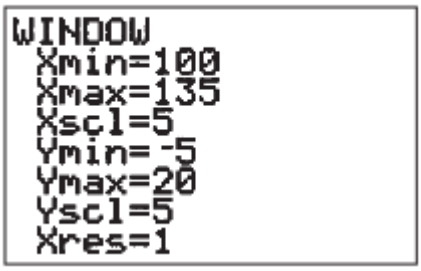
- Enter data as a column under L1.

Set WINDOW settings:

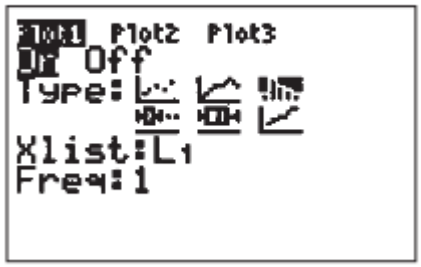
X min = 1 (minimum lower class boundary); X max = 41 (maximum upper class boundary);
X scl = 5 (for class width); Y min = -1; Y max = 30 (larger than max frequency value);
Y scl = 1. Note: the Xmin and Xmax values must allow for whole class widths to be displayed.

- Press [Y =] and clear any functions.
- Press [2nd] [Y=] for STAT PLOT.
- Choose plot 1 and turn it ON by choosing the word On and then press ENTER.
- Select the histogram symbol (press ENTER)
- select L1 for Xlist (source of data),
- Select L_2 for Freq (Press [2nd] 2)
- Press GRAPH to display the histogram

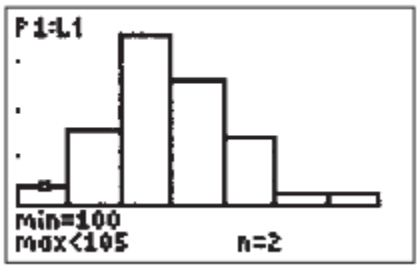
Input



Input



Output



Stem and Leaf Plots

The stem and leaf plot is a method of organizing data and is a combination of sorting and graphing. It has the advantage over a grouped frequency distribution of retaining the actual data while showing them in graphical form.

A **stem and leaf plot** is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes. Following example shows the procedure for constructing a stem and leaf plot.

Example 5: At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

Solution:

Step 1 Arrange the data in order:

02, 13, 14, 20, 23, 25, 31, 32, 32, 32, 32, 33, 36, 43, 44, 44, 45, 51, 52, 57

Step 2 Separate the data according to the first digit, as shown.

02
13, 14
20, 23, 25
31, 32, 32, 32, 32, 33, 36
43, 44, 44, 45
51, 52, 57

Step 3 A display can be made by using the leading digit as the stem and the trailing digit as the leaf. For example, for the value 32, the leading digit, 3, is the stem and the trailing digit, 2, is the leaf. For the value 14, the 1 is the stem and the 4 is the leaf. Now a plot can be constructed as shown below.

Leading digit (stem)	**Trailing digit (leaf)[One's place]**
0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

Example 6: The heights (to the nearest centimetre) of boys and girls in a grade 10 class in Bayview S.S are as follows:

165 171 169 169 172 171 171 180 168 168
166 165 171 173 187 181 175 174 165 167
163 160 169 167 172 174 177 188 177 185

Construct a stem-and-leaf display for these data.

Practice

1. The test score, out of 50 marks, is recorded for a group of 45 Geography students.

35 29 39 27 26 29 36 41 45 29 25 50 30 33 34
22 35 48 20 32 34 39 41 46 35 35 43 45 50 30
34 36 25 42 36 25 20 18 9 40 32 33 28 33 34

a) Construct an ordered stem-and-leaf plot for this data using 0, 1, 2, 3, 4 and 5 as the stems.
b) What advantage does a stem-and-leaf plot have over a frequency table?
c) What is the i) highest ii) lowest mark scored for the test?
d) If an 'A' was awarded to students who scored 42 or more for the test, what percentage of students scored an 'A'?
e) What percentage of students scored less than half marks for the test?

GDC ACTIVITY FOR HISTOGRAMS

Always start by clearing your RAM: [2] [+] [7] [1] [2] **→A RAM Cleared message should appear**

Raw Data Example:

The following figures are the heights (in centimetres) of a group of students:

156	172	168	153	170	160	170	156	160	160	172	174
150	160	163	152	157	158	162	154	159	163	157	160
153	154	152	155	150	150	152	152	154	151	151	154

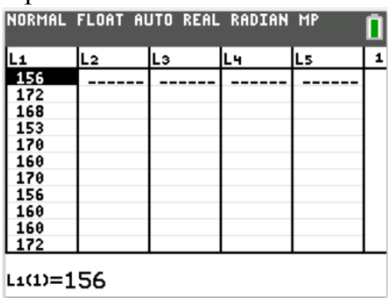
Use a GDC to construct a frequency histogram.

GDC Instructions For RAW DATA

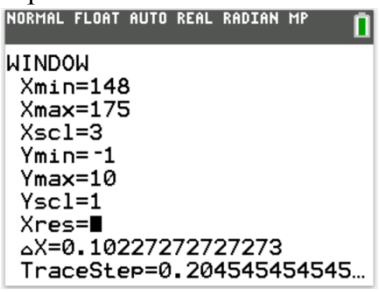
1) To enter the original data, press the [stat] key and choose [1] for EDIT from the screen menus. If necessary, press [stat] [4] followed by the keys [2nd] [1] (for L1) [,] [2nd] [2] (for L2), etc and then press [enter] to clear any previous lists.
2) Enter data as a column under L1. Press [enter] or the cursor buttons to move to the next cell until you are done entering all the values.
3) Set the WINDOW settings. Press [window], then enter the following values:
   X min = 148 (for minimum lower class boundary)
   X max = 175 (for maximum upper class boundary)
   X scl = 3 (for class width)
   Y min = -1 (equal to or smaller than the lowest frequency value)
   Y max = 10 (larger than max frequency value)
   Y scl = 1 (for the scale in the vertical axis)
   Note: the Xmin and Xmax values must allow for whole class widths to be displayed.
4) Press [2nd] [y=] for STAT PLOT. Choose plot [1] and turn it on by scrolling to the word On and then press [enter] to highlight it. Select the histogram symbol (press [enter]), select L1([2nd] [1]) for Xlist (for the source of data), select 1 for Freq (since each height is entered separately, the GDC will group them)
5) Press [graph] to display the histogram.

GDC Displays:

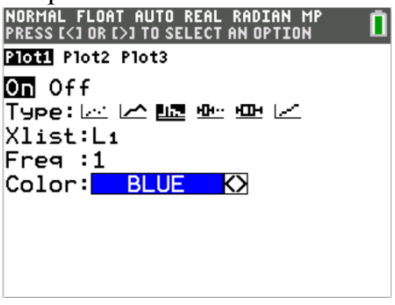
Steps 1 & 2:



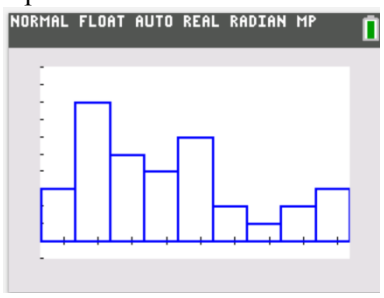
Step 3:



Step 4:



Step 5:



Grouped Data Example:

100 students take a mathematics test that has a maximum possible score of 40. The results are shown in the table. Use a GDC to construct a frequency histogram and a frequency density histogram.

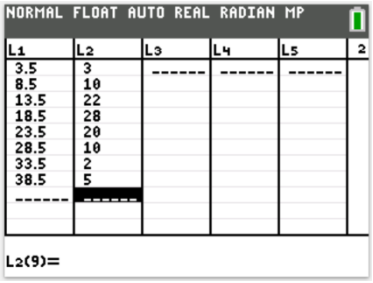
Mark	# students	Mid-point value	
$1 \le m < 6$	3		
$6 \le m < 11$	10		
$11 \le m < 16$	22		
$16 \le m < 21$	28		
$21 \le m < 26$	20		
$26 \le m < 31$	10		
$31 \le m < 36$	2		
$36 \le m < 41$	5		

GDC Instructions For GROUPED DATA

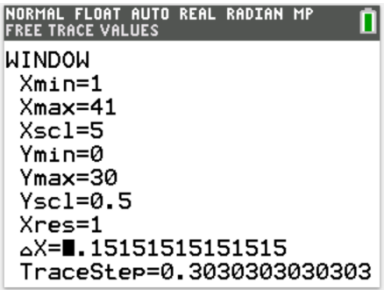
1) To enter the original data, press the [stat] key and choose [1] for EDIT from the screen menu. If necessary, press [stat] [4] followed by the keys [2nd] [1] (for L1) [,] [2nd] [2] (for L2), etc and then press [enter] to clear any previous lists.
2) Enter data. The **mid-point values of each class interval (Xlist) should be entered under L1 and the frequencies or frequency densities (Freq) under L2**. Press [enter] or the cursor buttons to move to the next cell until you are done entering all the values. Use the cursor buttons to move between L1 and L2.
3) Set the WINDOW settings. Press [window], then enter the following values:
   X min =1 (for minimum lower class boundary)
   X max = 41 (for maximum upper class boundary)
   X scl = 5 (for class width)
   Y min = 0 (equal to or smaller than the lowest frequency value)
   Y max = 30 (larger than max frequency value)
   Y scl = 0.5 (for the scale in the vertical axis)
   Note: the Xmin and Xmax values must allow for whole class widths to be displayed.
4) Press [2nd] [y=] for STAT PLOT. Choose plot [1] and turn it on by scrolling to the word On and then press [enter] to highlight it. Select the histogram symbol (press [enter]), select **L1([2nd] [1]) for Xlist** (for the source of the independent variable), Select **L2([2nd] [2]) for Freq** (since the frequencies are entered in column L2)
5) Press [graph] to display the histogram.

GDC Displays:

Steps 1 & 2:

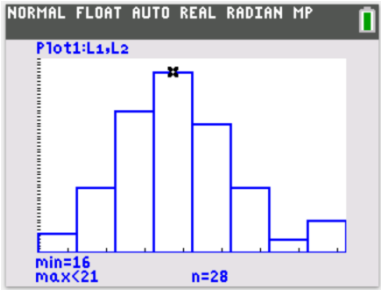


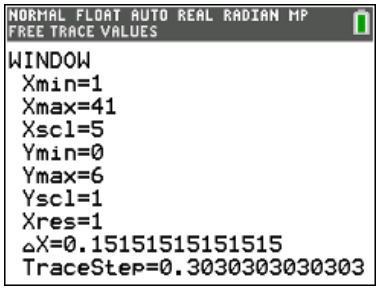
Step 3:

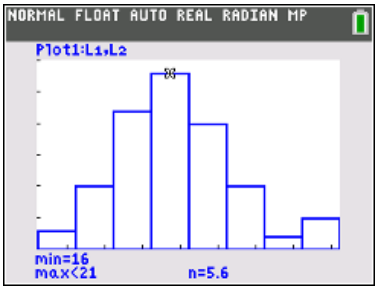


Step 4:

Step 5: Frequency Histogram or Frequency Density Histogram







Cumulative Frequency

The **cumulative frequency** is the total frequency up to a particular value or class boundary. It is calculated by 'accumulating' or adding all the previous frequencies as you move down the frequency table. The following example shows us how to construct a cumulative table and then draw a **cumulative frequency curve** that has effectively placed the data in order.

On a cumulative frequency curve, points plotted will be of the ordered pair: (upper class boundary, cumulative frequency). This point represents the number of data less than the endpoint of the interval.

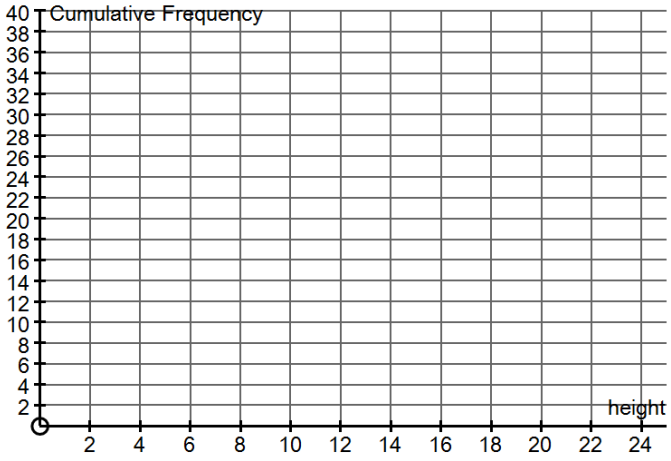
Example 1: The heights to the nearest centimetre of a type of plant were recorded 6 months after planting. The frequency distribution is shown in the table.

Height (cm)	Frequency
$3 \leq h < 6$	1
$6 \leq h < 9$	3
$9 \leq h < 12$	6
$12 \leq h < 15$	10
$15 \leq h < 18$	12
$18 \leq h < 21$	4

Show these results on a **cumulative frequency curve**.

Height (cm)	Frequency	Upper Class Boundary	Cumulative Frequency
$3 \leq h < 6$	1		
$6 \leq h < 9$	3		
$9 \leq h < 12$	6		
$12 \leq h < 15$	10		
$15 \leq h < 18$	12		
$18 \leq h < 21$	4		

Answer:



What does the point $(18, 32)$ on the cumulative frequency curve tell you about the data?

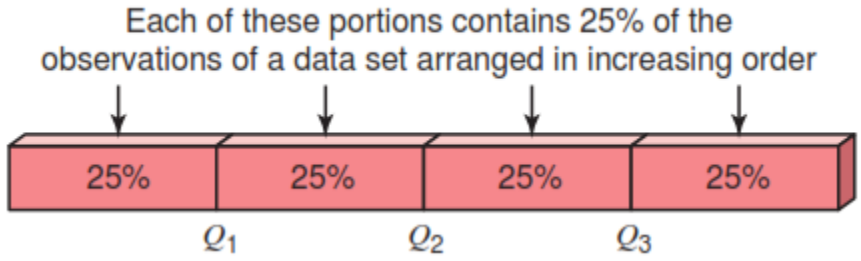
Measure of Position

A measure of position determines the position of a single value in relation to other values in a sample or a population data set.

Quartiles and Interquartile Range
Quartiles are the summary measures that divide a ranked data set into four equal parts. The quartiles are defined as follows.

Definition

- The **median**, Q_2**,** Second quartile given by the value of the middle term in a ranked data set
- The **first quartile,** Q_1, given by the value of the middle term among the (ranked) observations that are less than the median
- The **third quartile,** Q_3, given by the value of the middle term among the (ranked) observations that are greater than the median



Calculating Interquartile Range

The difference between the third and the first quartiles gives the **interquartile range**; that is, IQR =Interquartile range = $Q_3 - Q_1$

The Quartiles can be found with the following formulas and rounding rules:

$$Q_1 = \begin{cases} \left(\frac{n+1}{4}\right)\text{th datum} & \text{if } n \text{ is odd} \\ \left(\frac{n+2}{4}\right)\text{th datum} & \text{if } n \text{ is even} \end{cases} \qquad Q_2 = 2\left(\frac{n+1}{4}\right)\text{th datum} \qquad Q_3 = \begin{cases} 3\left(\frac{n+1}{4}\right)\text{th datum} & \text{if } n \text{ is odd} \\ \left(\frac{3n+2}{4}\right)\text{th datum} & \text{if } n \text{ is even} \end{cases}$$

- Since $n \epsilon N$ and each formula has a division by 4, the only possible decimals you can get are:
  - 0: In this case, take the datum value.
  - 0.25: In this case, round down, and take the datum value.
  - 0.5: In this case, take the rounded down datum value and the rounded up datum value and take the mean.
  - 0.75: In this case, round up, and take the datum value.

Example 1: Determine the median and the quartiles for the following data: 4, 16, 59, 77, 88, 93.

Example 2: The 2008 profits (rounded to billions of dollars) of 12 companies selected from all over the world is reproduced below.

Company	2008 Profits (billions of dollars)
Merck & Co	8
IBM	12
Unilever	7
Microsoft	17
Petrobras	14
Exxon Mobil	45
Lukoil	10
AT&T	13
Nestlé	17
Vodafone	13
Deutsche Bank	9
China Mobile	11

(a) Find the values of the three quartiles. Where does the 2008 profits of Merck & Co fall in relation to these quartiles?
(b) Find the interquartile range.

Solution:

Example 3: The following are the ages (in years) of nine employees of an insurance company:

47 28 39 51 33 37 59 24 33

(a) Find the values of the three quartiles. Where does the age of 28 years fall in relation to the ages of these employees?
(b) Find the interquartile range.

GDC INSTRUCTIONS FOR FINDING INTERQUARTILE RANGE

Sample problem: Find the TI 83 interquartile range for the for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars.

11.2 11.9 12.0 12.8 13.4 14.3

Step 1: Press the STAT button and then press ENTER. Enter the first number (11.2), and then press ENTER. Continue entering numbers, pressing ENTER after each entry.
Step 2: Press the STAT button.
Step 3: Press the right arrow button to select "Calc."
Step 4: Press ENTER to highlight "1-Var Stats."
Step 5: Press ENTER again to bring up a list of stats.
Step 6: Scroll down the list with the arrow keys to find Q1 and Q3. Write those numbers down.
Step 7: Subtract Q1 from Q3 to find the **IQR**

Output

```
1-Var Stats
 x̄=12.6
 Σx=75.6
 Σx²=958.94
 Sx=1.1296017
 σx=1.031180553
↓n=6
```

Output

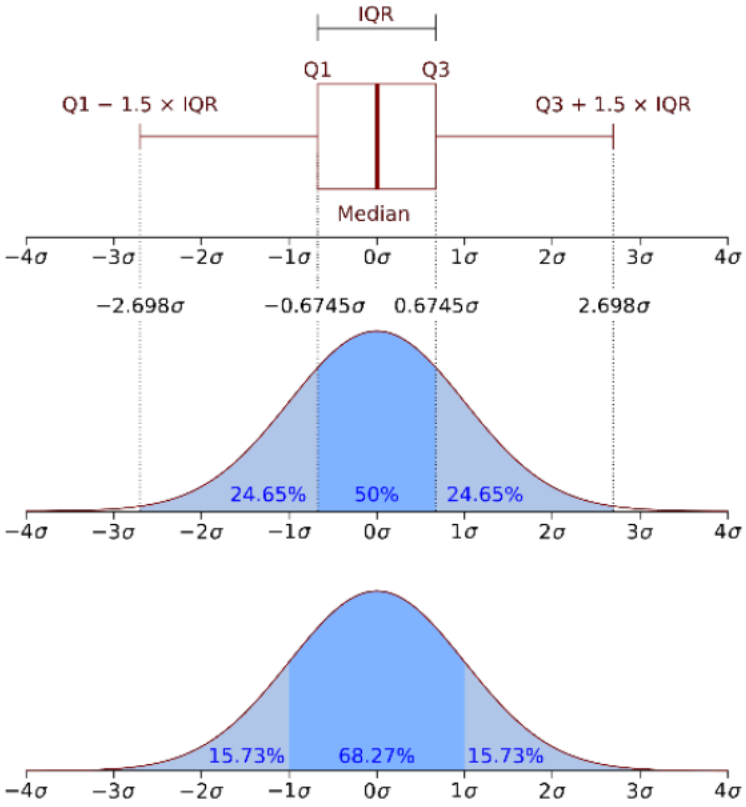
```
1-Var Stats
↑n=6
 minX=11.2
 Q1=11.9
 Med=12.4
 Q3=13.4
 maxX=14.3
```

Lower Inner fence and Upper Inner fence

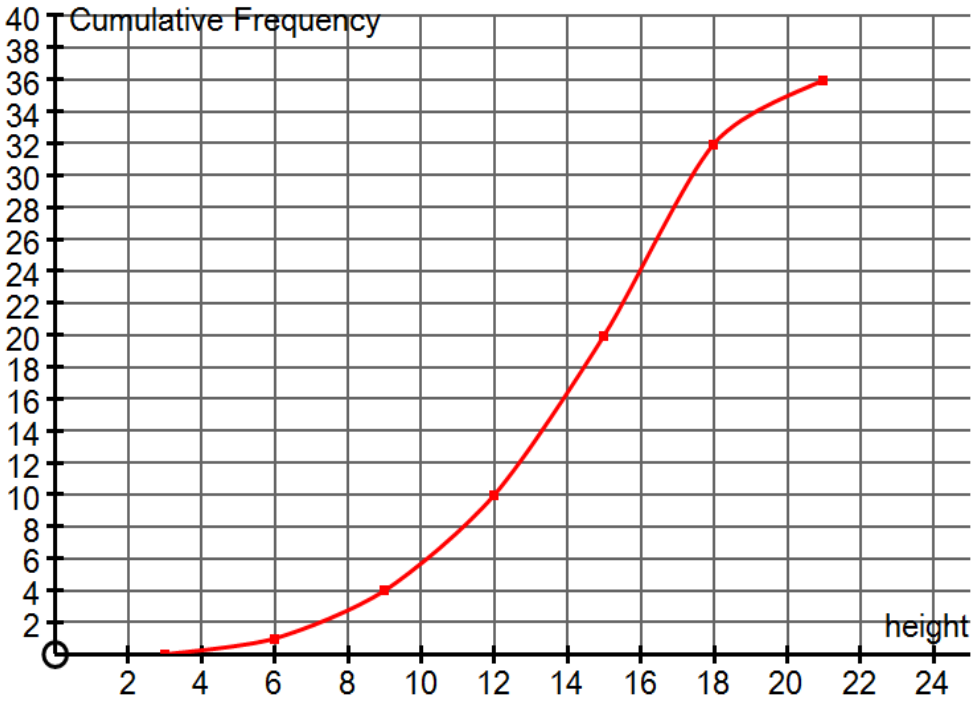
As a rule of thumb, data values that deviate from the middle value by more than twice the IQR deserve individual attention. They are called "outliers." Data values that deviate from the middle value by more than 3.5 times the IQR are usually scrutinized closely. They are sometimes called "far outliers."
There are many methods to check for **outliers** in Statistics. One such method is to calculate the Lower Inner fence and Upper Inner fence.

- Lower Inner fence $= Q_1 - 1.5(\text{IQR})$
- Upper Inner fence $= Q_3 + 1.5(\text{IQR})$

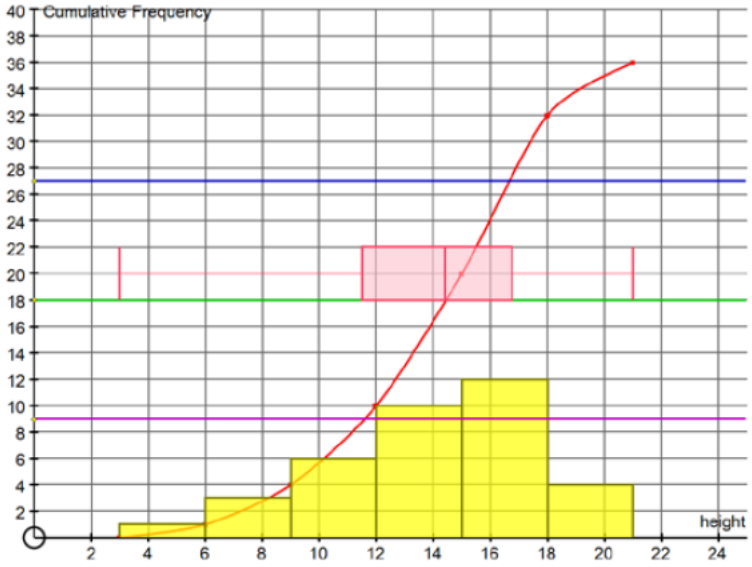
If data points are outside of the interval, [Lower Inner fence, Upper Inner fence], they may be considered to be **outliers**. Note that fences only provide a guideline by which to define an outlier.



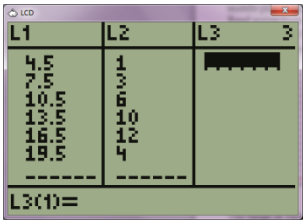
Example 4: Determine the median, quartiles, and IQR from the cumulative frequency graph given below.
Note: To start, note that there are 36 observations and so $n = 36$ for your quartile calculations.

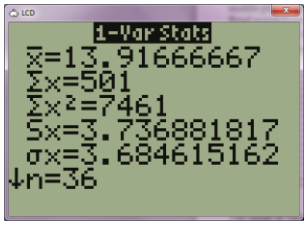


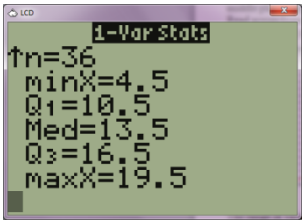
Solution diagram:



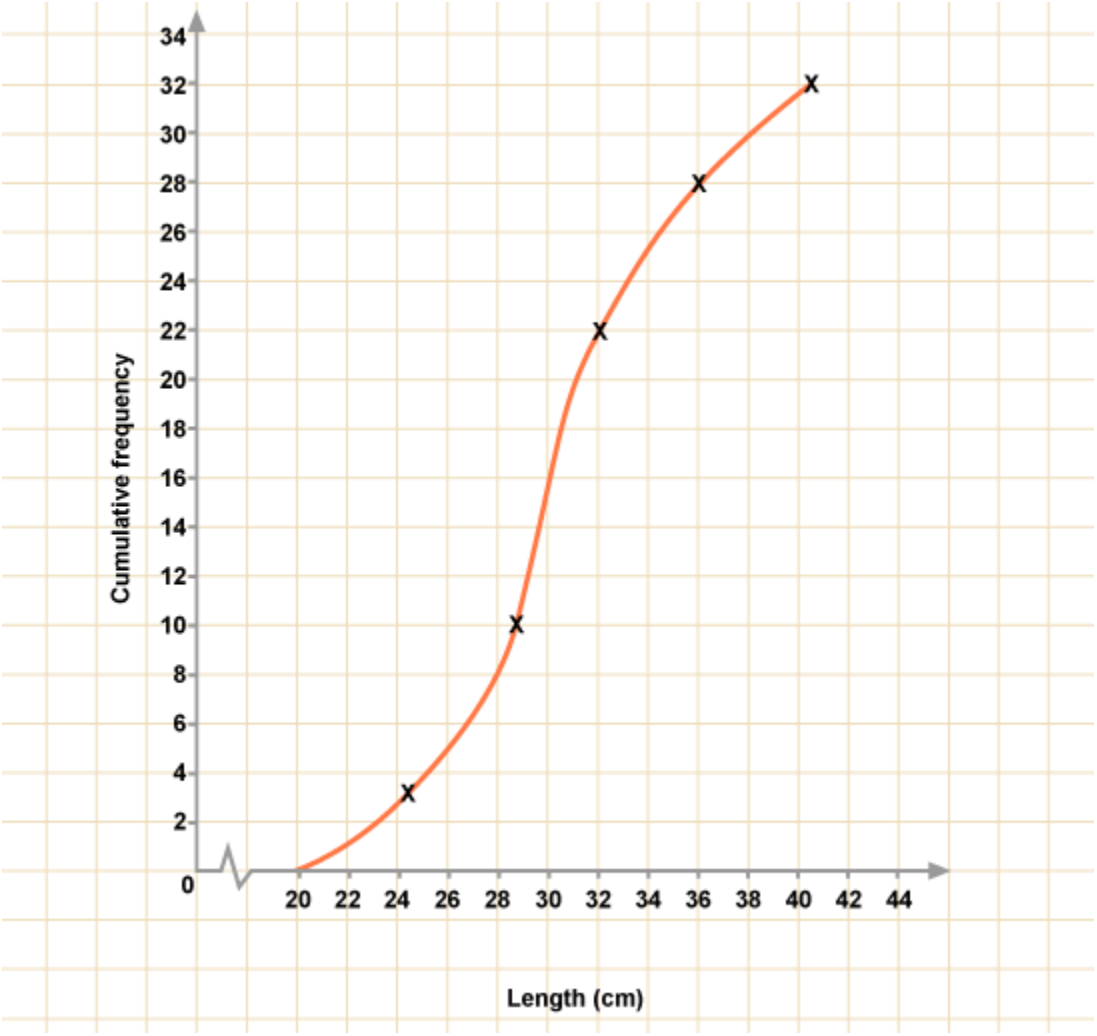
Note: When calculating quartiles from the table, this may yield different results than determining the quartiles from the cumulative frequency graph due to rounding that must be done to calculate the mid-values (See L1 below).





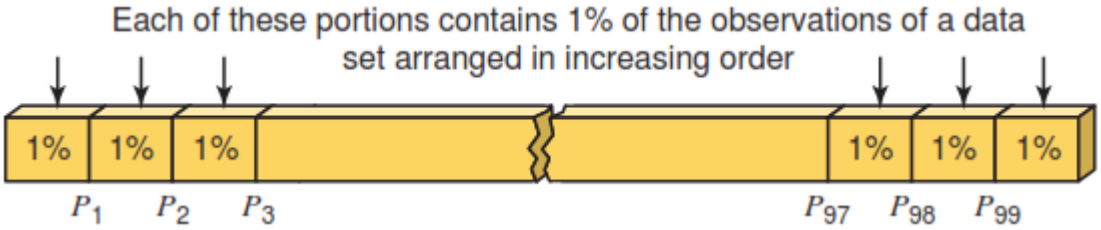


Example 5: Determine the median, quartiles, and IQR from the cumulative frequency graph given below.



Percentiles

Percentiles are the summary measures that divide a ranked data set into 100 equal parts. Each (ranked) data set has 99 percentiles that divide it into 100 equal parts. The data should be ranked in increasing order to compute percentiles. The kth percentile is denoted by P_k, where k is an integer in the range 1 to 99. For instance, the 25th percentile is denoted by P_{25}.



The approximate value of the k^{th} percentile is determined as explained next.

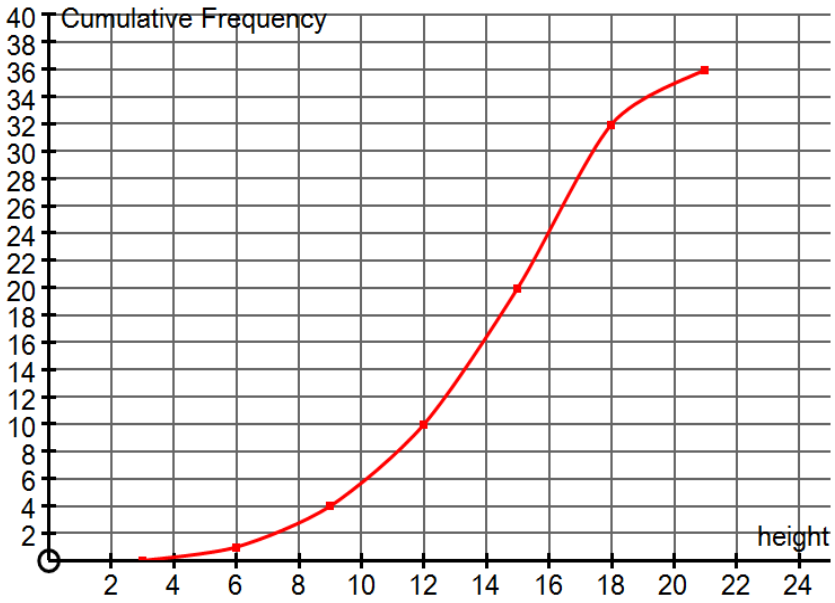
Calculating Percentiles

The (approximate) value of the k^{th} percentile, denoted by P_k , is

$$P_k = \text{Value of the } \left(\frac{kn}{100}\right) \text{th term in a ranked data set}$$

where k denotes the number of the percentile and n represents the sample size.

Example 6: Determine the 30^{th} and 70^{th} percentile using the cumulative frequency curve given in example 4.



Cumulative Frequency Polygon

Instead of forming a curve, you can join the points with straight lines.

Example 7: In a cricket match the 40 completed innings gave this distribution of scores.

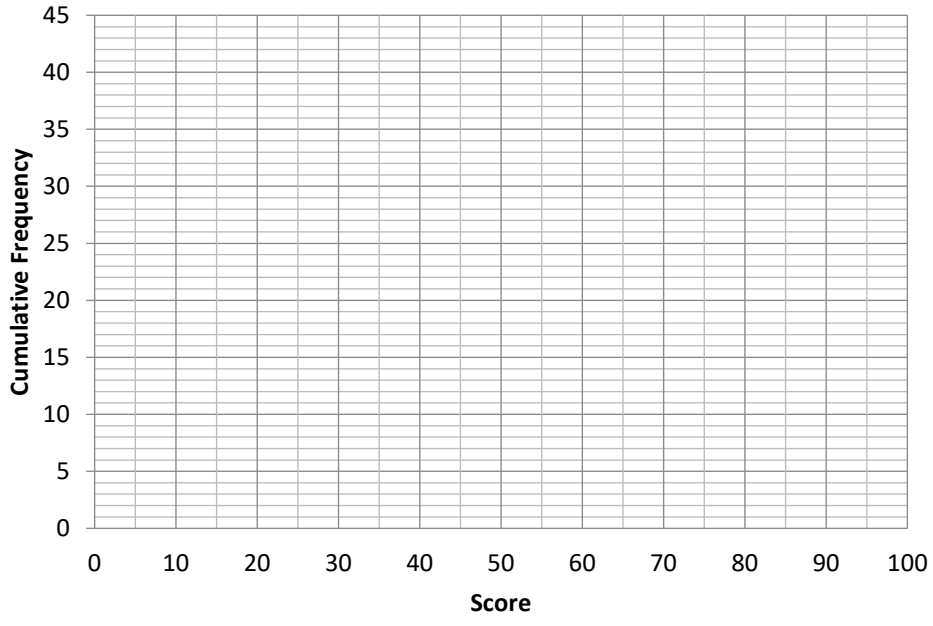
Score	$0 \leq s < 10$	$10 \leq s < 20$	$20 \leq s < 30$	$30 \leq s < 40$	$40 \leq s < 50$	$50 \leq s < 60$	$60 \leq s < 70$	$70 \leq s < 80$	$80 \leq s < 90$	$90 \leq s < 100$
Freq.	8	10	6	5	6	2	0	2	0	1

a) Show these results on a cumulative frequency polygon.
b) Use your polygon to estimate the median score.
c) From your polygon, find the upper and lower quartiles, and hence, estimate the interquartile range.
d) The cricket club decide to award prizes to the top 20% of players in the match. What score should be used as a minimum to award prizes?

Solution:

Score	**Upper boundary**	**Frequency**	**Cumulative Frequency**
$0 \leq s < 10$		8	
$10 \leq s < 20$		10	
$20 \leq s < 30$		6	
$30 \leq s < 40$		5	
$40 \leq s < 50$		6	
$50 \leq s < 60$		2	
$60 \leq s < 70$		0	
$70 \leq s < 80$		2	
$80 \leq s < 90$		0	
$90 \leq s < 100$		1	

a) Cumulative frequency polygon.

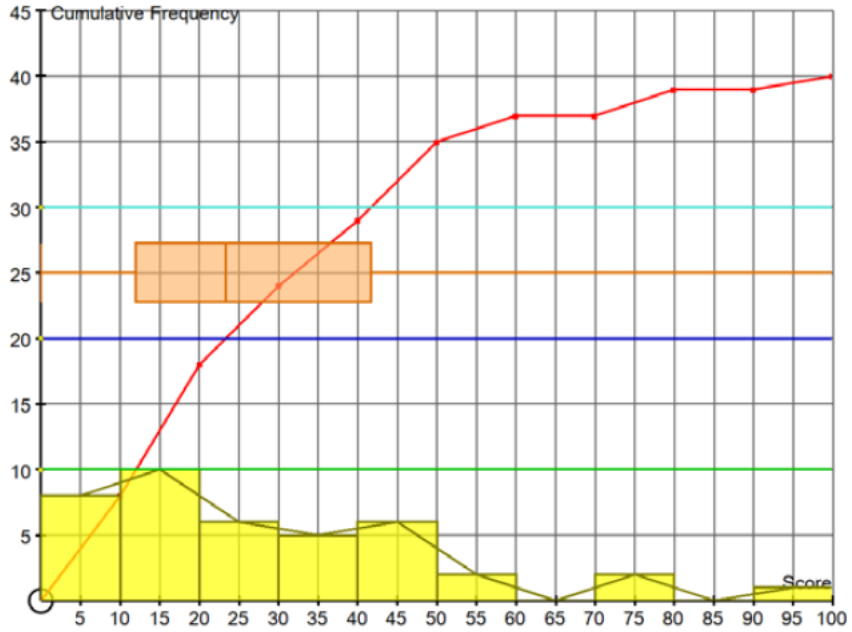


b) Use your polygon to estimate the median score.

c) From your polygon, find the upper and lower quartiles, and hence, estimate the interquartile range.

d) The cricket club decide to award prizes to the top 20% of players in the match. What score should be used as a minimum to award prizes? (Hint: Calculate P_{80})

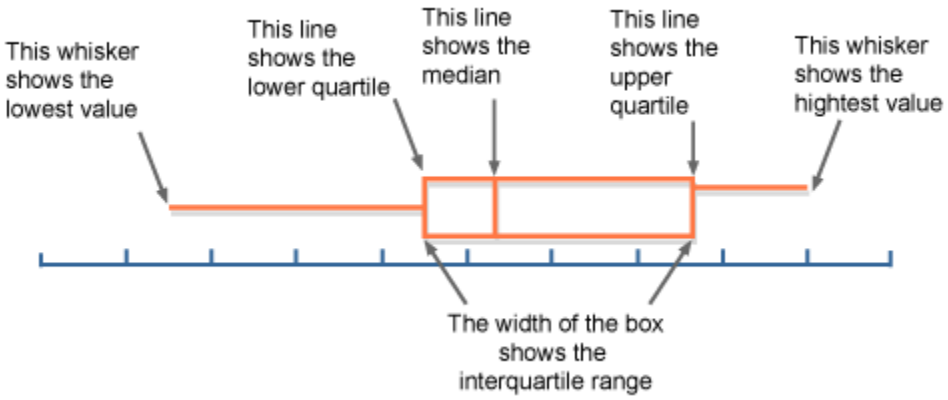
Solution Diagram



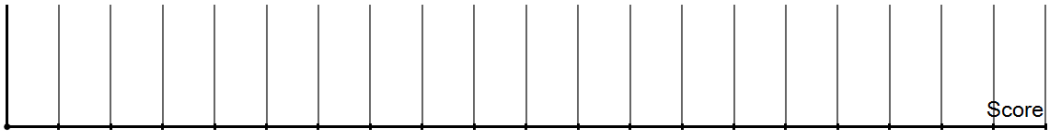
Box and Whiskers Plots

A box-and-whisker plot gives a graphic presentation of data using five measures: the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.
A box-and-whisker plot can help us visualize the center, the spread, and the skewness of a data set. It also helps detect outliers. We can compare different distributions by making box-and-whisker plots for each of them.

A **box and whiskers plot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q1 , drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q1 and Q3 with a vertical line inside the box passing through the median or Q2.

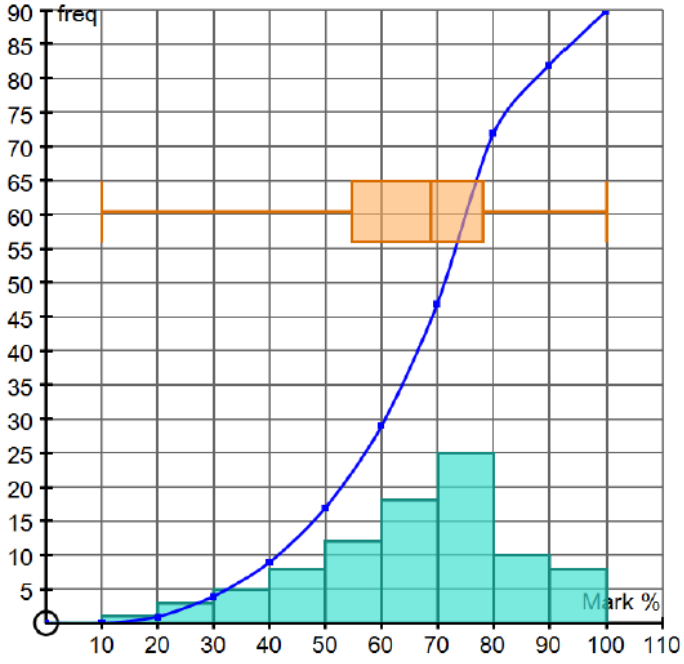


Example 8: Illustrate the data from Example 7 using a box and whisker plot.



Example 9: The following data shows the percentage scores from a Shakespeare English Test from 3 different classes in the year of 1996.

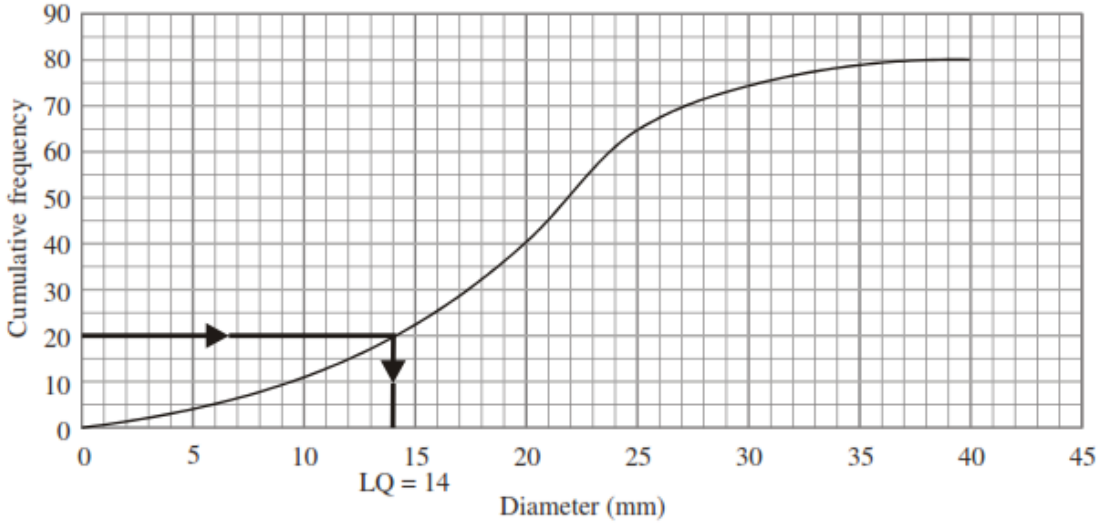
Score	Upper boundary	Frequency	Cumulative Frequency
$0 \leq s < 10$		0	
$10 \leq s < 20$		1	
$20 \leq s < 30$		3	
$30 \leq s < 40$		5	
$40 \leq s < 50$		8	
$50 \leq s < 60$		12	
$60 \leq s < 70$		18	
$70 \leq s < 80$		25	
$80 \leq s < 90$		10	
$90 \leq s < 100$		8	



a) Using a sheet of lined paper or grid paper, show the results on a cumulative frequency curve.
b) Use your curve to estimate Q1, Q2, Q3, and IQR. [54.5%, 68.9%, 78.2%, 23.6%]
c) The 1 student who got in the 10% - 20% range might be an outlier. Justify why.
d) Construct a box and whisker plot using the results from part b).
e) If a sticker is to be placed on the assessment of the top 10% of students, what score should be used as the minimum? [around 88%]

Exit Card!

1) A student measured the diameters of 80 snail shells. His results are shown in the following cumulative frequency graph. The lower quartile (LQ) is 14 mm and is marked clearly on the graph.



(a) On the graph, mark clearly in the same way and write down the value of

(i) the median;
(ii) the upper quartile.

(b) Write down the interquartile range.

2) A set of data is 18, 18, 19, 19, 20, 22, 22, 23, 27, 28, 28, 31, 34, 34, 36. The box and whisker plot for this data is shown below.

(a) Write down the values of A, B, C, D and E.

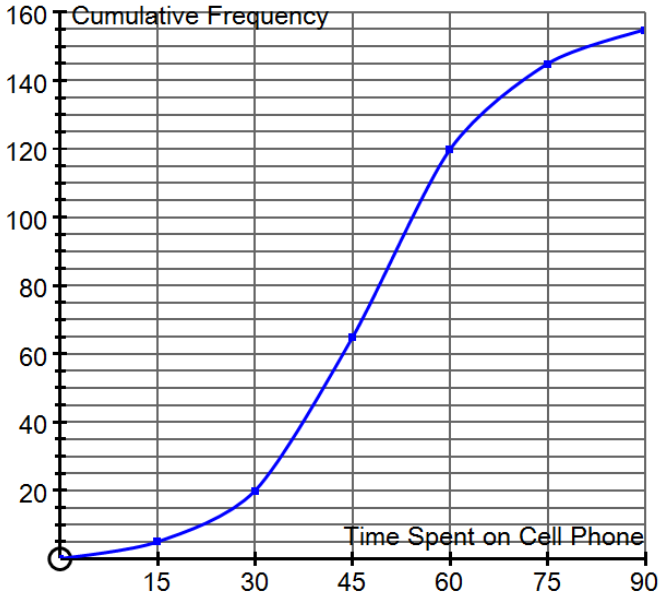
A = ... B = ... C= ..... D = ... E = ...

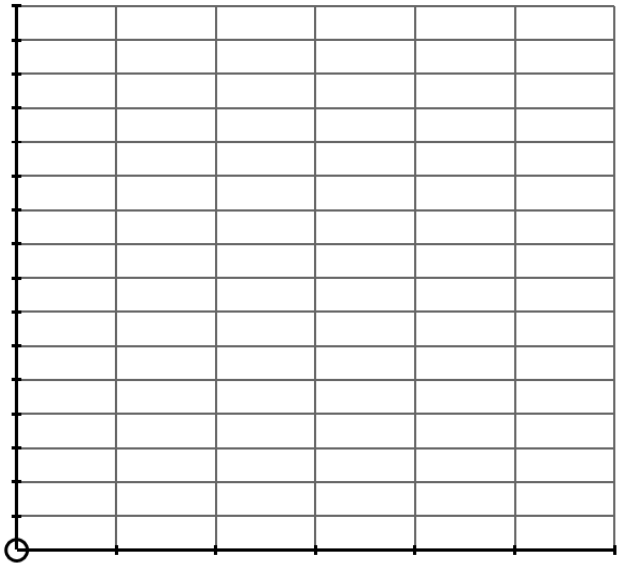
(b) Find the interquartile range.

Warm Up

1. The Cumulative Frequency Graph shows the results of a survey which sampled individual Cell Phone use for the day.

   a. Create a labeled histogram of the dataset on the grid provided.





   b. What percentage of people spend more than 60 minutes on their Cell Phones a day?

2. In a school with 125 girls, each student is tested to see how many sit-up exercises (sit-ups) she can do in one minute. The results are given in the table below.

Number of sit-ups	**Number of students**	**Cumulative number of students**
15	11	11
16	21	32
17	33	***p***
18	***q***	99
19	18	117
20	8	125

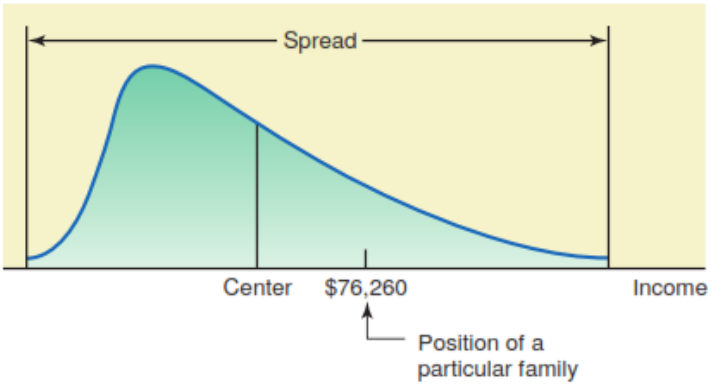
(a) Write down the value of p.

(b) Find the value of q.

Numerical Descriptive Measures

In previous chapter we discussed how to summarize data using different methods and to display data using graphs. Graphs are one important component of statistics; however, it is also important to numerically describe the main characteristics of a data set. The numerical summary measures, such as the ones that identify the center and spread of a distribution, identify many important features of a distribution. For example, the techniques we've learned so far can help us graph data on family incomes. However, if we want to know the income of a "typical" family (given by the center of the distribution), the spread of the distribution of incomes, or the relative position of a family with a particular income, the numerical summary measures can provide more detailed information (See below figure). The measures that we discuss in this unit include

(1) Measures of central tendency
(2) Dispersion (or spread)
(3) Position



1. Measure of Tendency

a) Measure of Tendency of Ungrouped Data

A measure of central tendency gives the center of a histogram or a frequency distribution curve. This section discusses three different measures of central tendency: the **mean**, the **median**, and the **mode**. We will learn how to calculate each of these measures for ungrouped data.

i) Mean

The mean, also called the arithmetic mean, is the most frequently used measure of central tendency. For ungrouped data, the mean is obtained by dividing the sum of all values by the number of values in the data set. To show a sum of the total X values, the symbol $\sum$ (the capital Greek letter sigma) is used, and $\sum X$ means to find the sum of the X values in the data set.

*The **mean** is the sum of the values, divided by the total number of values. The symbol represents the **sample mean** $\bar{x}$.*

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where **n** represents the total number of values in **the sample**.

Rounding Rule for the Mean

In statistics the basic rounding rule is that when computations are done in the calculation, rounding should not be done until the final answer is calculated. When rounding is done in the intermediate steps, it tends to increase the difference between that answer and the exact one. But in the textbook

and solutions manual, it is not practical to show long decimals in the intermediate calculations; hence, the values in the examples are carried out to enough places (**usually three or four**) to obtain the same answer that a calculator would give after rounding on the last step.

Example 1: The data represent the number of days off per year for a sample of individuals selected from nine different countries. Find the mean.

20, 26, 40, 36, 23, 42, 35, 24, 30

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A	B	C	D
Class	**Frequency (*fi*)**		

Step 2 Find the midpoints of each class and place them in column C.
Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.
Step 4 Find the sum of column D.
Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is $\bar{x} = \frac{\sum \mathbf{x_i f_i}}{\sum \mathbf{f_i}}$

Example 2: The following is a **frequency table** of the heights (in centimetres) of a group of students. Find an estimate of the mean height of the students

Heights (cm)	**Mid-height(x_i)**	**# of students** f_i	$f_i \times x_i$
$148 \le h < 151$	149.5	3	
$151 \le h < 154$		8	
$154 \le h < 157$		7	
$157 \le h < 160$		4	
$160 \le h < 163$		6	
$163 \le h < 166$		2	
$166 \le h < 169$		1	
$169 \le h < 172$		2	
$172 \le h < 175$		3	
	Totals	$\sum f_i =$	$\sum f_i \cdot x_i =$

ii) The Median

The **median** is the *middle value* when all the data are arranged in order of magnitude.

- If the number of data, *n,* is odd then the median is the value of the $\frac{n+1}{2}$ term in the sorted list.
- If *n* is even, then the median is the mean of the two middle data. The median is the mean of the $\frac{n}{2}$ and $(\frac{n}{2}+1)$ term in the sorted list.

Example 3: Find the median of the following sets of data.
a) In an office block, the amount (to the nearest euro) spent on lunch by a cross-section of office workers on a particular Friday was recorded as: 4, 14, 2, 6, 6, 4, 24, 10, 12

b) The ages of university students in a tutorial group were recorded as: 20, 23, 18, 19, 28, 26, 22, 18

iii) The Mode

The third measure of average is called the *mode*. The mode is the value that occurs most often in the data set. It is sometimes said to be the most typical case.

The value that occurs most often in a data set is called the ***mode.***

A data set that has only one value that occurs with the greatest frequency is said to be **unimodal**. If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal**. If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal**. When no data value occurs more than once, the data set is said to have **no mode**. A data set can have more than one mode or no mode at all. These situations will be shown in some of the examples that follow.

Example 4: The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode. *Source: The World Almanac and Book of Facts.*

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

Example 5: Find the modal class for the frequency distribution of miles that 20 runners ran in one Week listed as follows.

Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

Properties and Uses of Central Tendency

The Mean

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

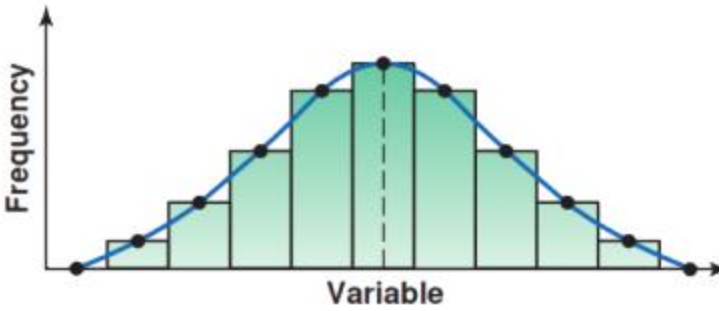
The Median

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for an open-ended distribution.
4. The median is affected less than the mean by extremely high or extremely low values.

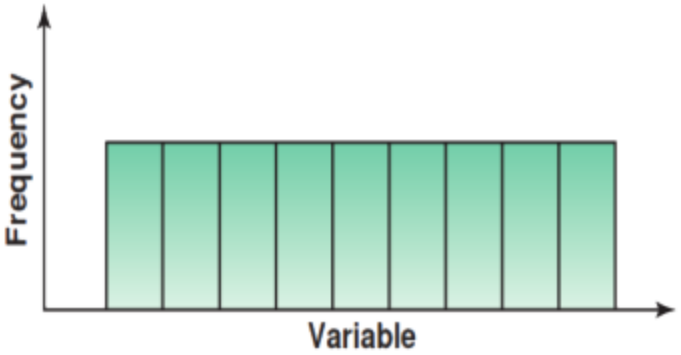
The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal, such as religious preference, gender, or political affiliation.
4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

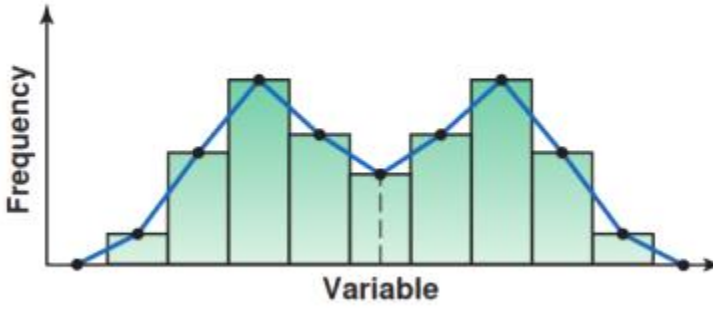
Distribution Shapes



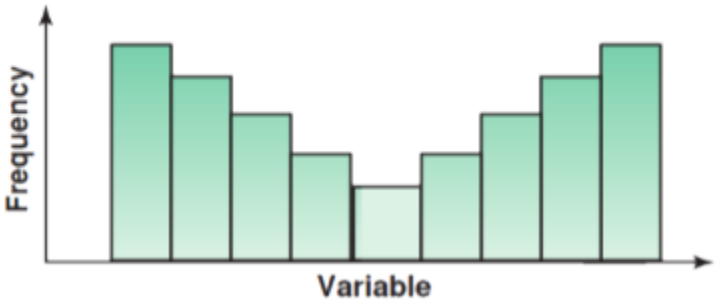
(a) Bell-shaped



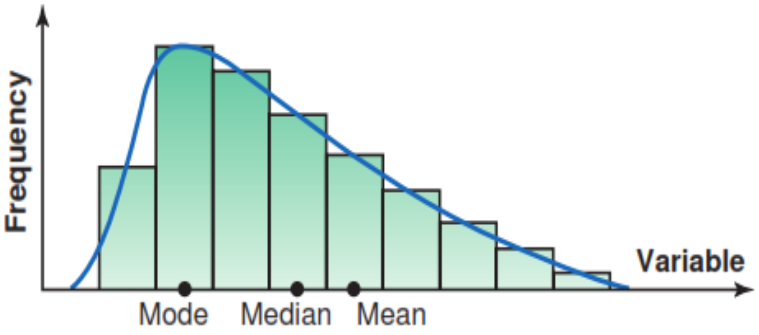
(b) Uniform



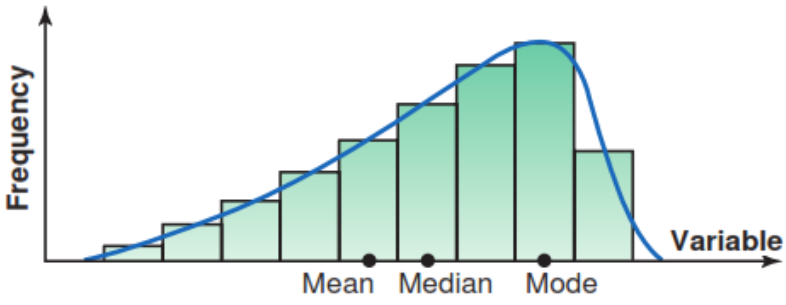
(c) Bimodal



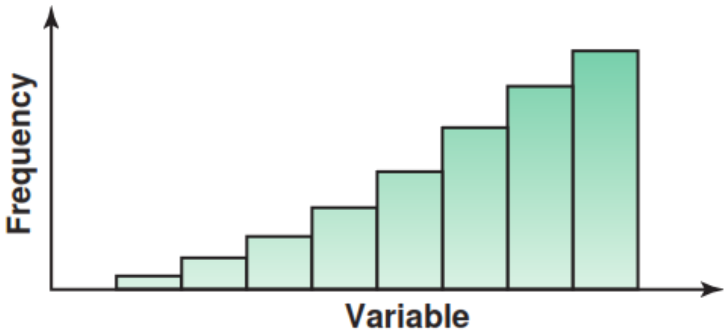
(d) U-Shaped



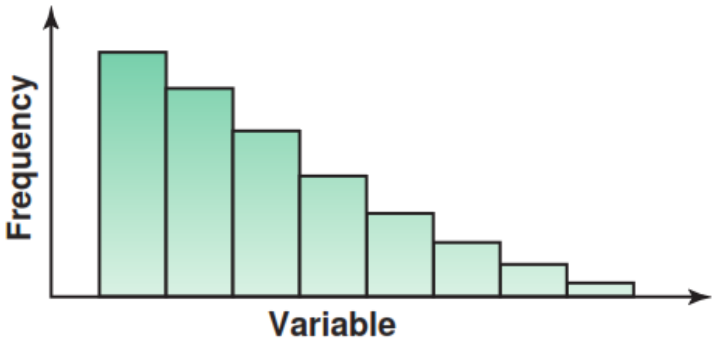
(e) Right-skewed



(f) Left-skewed



(g) J-shaped



(h) Reverse J-shaped

Practice

1. Describe which measure of central tendency—mean, median, or mode—was probably used in each situation.

(a) One-half of the factory workers make more than $5.37 per hour, and one-half make less than $5.37 per hour.
(b) The average number of children per family in the Plaza Heights Complex is 1.8.
(c) Most people prefer red convertibles over any other color.
(d) The average person cuts the lawn once a week.
(e) The most common fear today is fear of speaking in public.
(f) The average age of university professors is 42.3 years.

2. If the mean of five values is 64, find the sum of the values.

3. If the mean of five values is 8.2 and four of the values are 6, 10, 7, and 12, find the fifth value.

4. Find the mean of 10, 20, 30, 40, and 50.

(a) Add 10 to each value and find the mean.
(b) Subtract 10 from each value and find the mean.
(c) Multiply each value by 10 and find the mean.
(d) Divide each value by 10 and find the mean.
(e) Make a general statement about each situation.

5. Twenty business majors and 18 economics majors go bowling. Each student bowls one game. The Score keeper announces that the mean score for the 18 economics majors is 144 and the mean score for the entire group of 38 students is 150. Find the mean score for the 20 business majors.

6. The mean income for five families in 2015 was $99,520. What was the total income of these five families in 2015?

7. The mean age of six persons is 46 years. The ages of five of these six persons are 57, 39, 44, 51, and 37 years, respectively. Find the age of the sixth person.

8. Melissa's grade in her math class is determined by three 100-point tests and a 200-point final exam. To determine the grade for a student in this class, the instructor will add the four scores together and divide this sum by 5 to obtain a percentage. This percentage must be at least 80 for a grade of B. If Melissa's three test scores are 75, 69, and 87, what is the minimum score she needs on the final exam to obtain a B grade?

Answer:

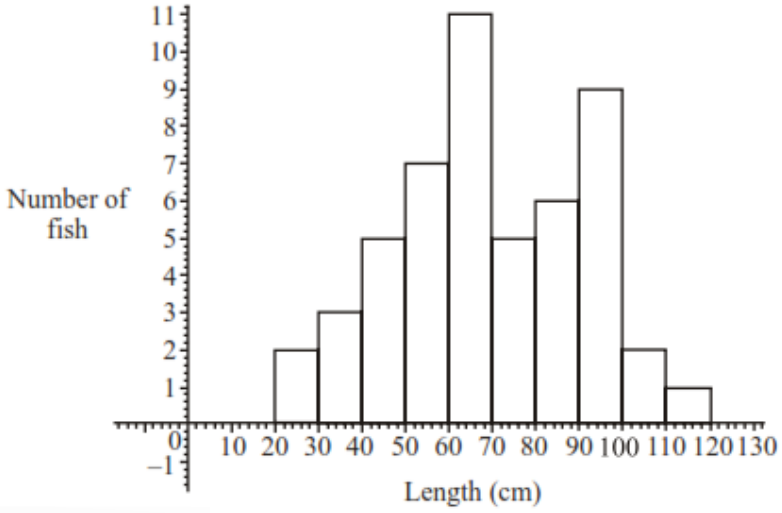
1. a) median b) mean c) mode d) mode e)mode f) mean
2. 320 3. 6 4. a) 40 b) 20 c) 300 d) 3
5. 155.4 6.$497,600 7.48 8. ≥ 169

Warm up

1. For what value of x the set of data 70, 110, x, 80, 60 will have the same median, mean and mode?

2. Seven consecutive whole numbers add up to 147.What is the result when their mean is subtracted from their median?

3. The figure below shows the lengths in centimeters of fish found in the net of a small trawler



(a) Find the total number of fish in the net.

(b) Find

i. the modal length interval;
ii. the interval containing the median length;
iii. an estimate of the mean length.

4. The heights of 200 students are recorded in the following table.

Height (h) in cm	Frequency
$140 \leq h < 150$	2
$150 \leq h < 160$	28
$160 \leq h < 170$	63
$170 \leq h < 180$	74
$180 \leq h < 190$	20
$190 \leq h < 200$	11
$200 \leq h < 210$	2

(a) Write down the modal group.

(b) Calculate an estimate of the mean of the heights.

Measures of Dispersion

In statistics, to describe the data set accurately, statisticians must know more than the measures of central tendency. Consider the following example.

Example 1: A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

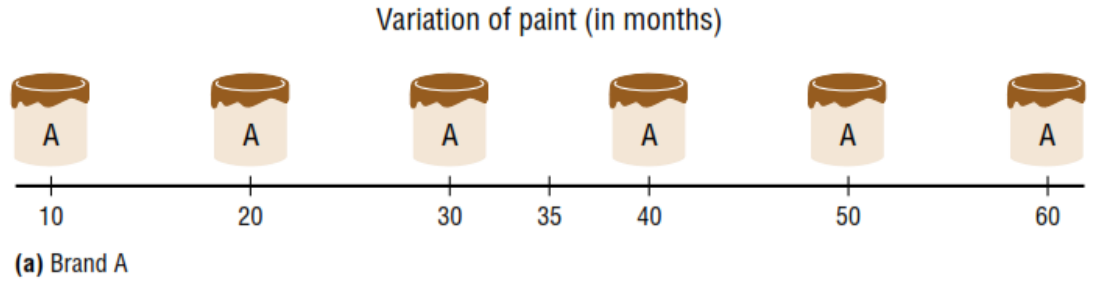
Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Solution:

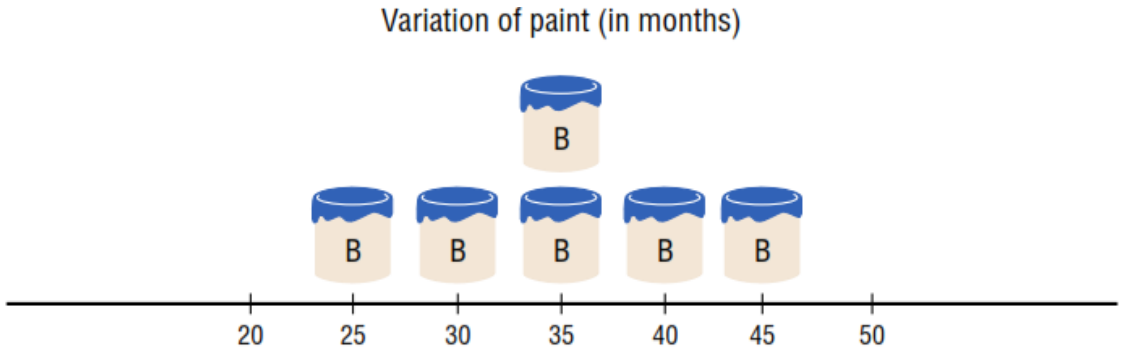
The mean for brand A is $\mu = \frac{210}{6} = 35$

The mean for brand B is $\mu = \frac{210}{6} = 35$

Since the means are equal, you might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. As below figure shows, even though the means are the same for both brands, the spread, or variation, is quite different. Figure shows that brand B performs more consistently; it is less variable. For the spread or variability of a data set, three measures are commonly used: ***range***, ***variance***, and ***standard deviation***. Each measure will be discussed in this section.



(a) Brand A



(b) Brand B

i) Range

The range is the simplest of the three measures and is defined now.

The range is the highest value minus the lowest value. The symbol R is used for the range.

R = highest value - lowest value

Example 2: The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.

Staff	Salary
Owner	$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
Workers	15,000
Workers	18,000

One extremely high or one extremely low data value can affect the range markedly, since the owner's salary is included in the data, the range is a large number. To have a more meaningful statistic to measure the variability, statisticians use measures called the *variance* and *standard deviation*.

ii) Variance and Standard Deviation

The standard deviation is the most-used measure of dispersion. The value of the standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a large value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean. The standard deviation is obtained by taking the positive square root of the variance. The variance calculated for population data is denoted by σ^2 (read as sigma squared). Consequently, the standard deviation calculated for population data is denoted by σ. Following is what we will call the basic formula that is used to calculate the variance.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} f_i (x_i - \mu)^2 \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{n}}$$

or

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i x_i^{\,2}}{n} - \mu^2 \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^{\,2}}{n} - \mu^2}$$

where k is the number of distinct values or classes, x_i is the midpoint and f_i is the frequency of a class and $n = \sum_{i=1}^{k} f_i$.

The quantity $x_i - \mu$ in the above formula is called the ***deviation*** of the x value from the mean.

The sum of the deviations of the x values from the mean is always zero; that is, $\sum_{i=1}^{k}(x_i - \mu) = 0$.

Example 3: The heights (in cm) of eight plants are measured 3 months after they are fed with a new plant food. Calculate the mean and standard deviation of the heights:
30, 17, 32, 25, 31, 28, 35, 26

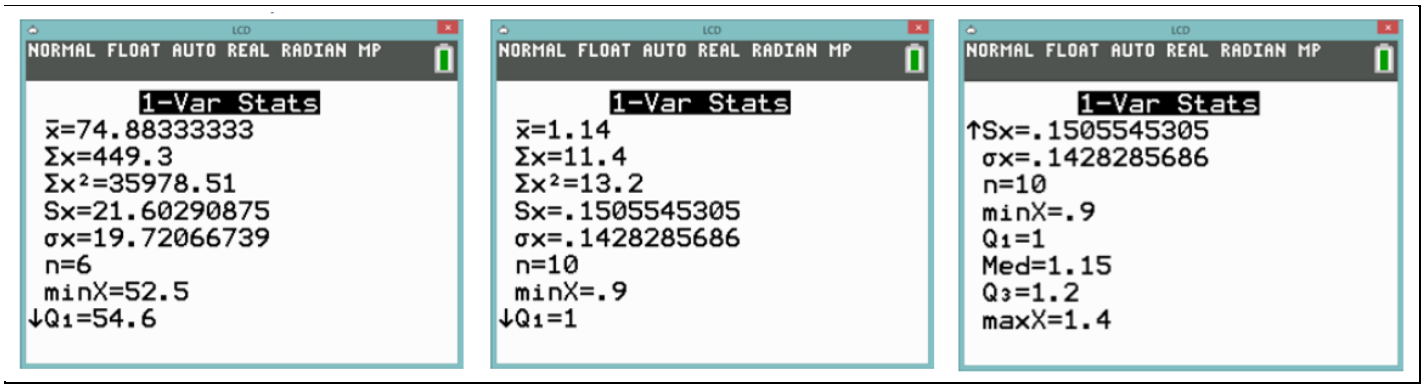
Example 4: The heights, in metres, of 10 children of a particular age in a school were recorded as:

1.0, 1.2, 1.3, 1.1, 1.2, 1.4, 1.1, 0.9, 1.3, 1.2

Calculate a) the mean b) the standard deviation. Compare the results with a graphics calculator.

GDC INSTRUCTIONS FOR 1-VARIABLE STATISTICS:

- Press STAT
- Press 1 (for EDIT) and ENTER
- Enter the data in L1
- Press STAT
- Select the CALC menu
- Select 1 (for 1-VAR Stats)
- Press 2nd 1 (for L1 which contains the data) followed by ENTER



Example 5: Find the variance and the standard deviation for the frequency distribution of the data Listed below. The data represent the number of miles that 20 runners ran during one week.

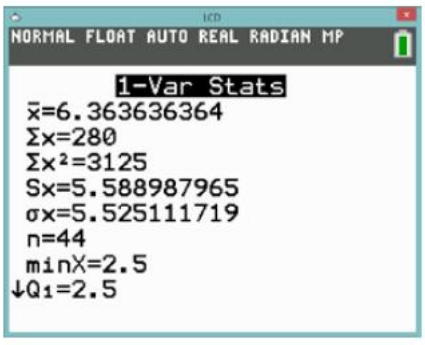
Distance (mi)	**Midpoint (x_i)**	**Frequency (f_i)**	x_i^2	$f_i.x_i$	$f_i.x_i^2$
$5 \le d < 11$	8	1			
$11 \le d < 16$	13	2			
$16 \le d < 21$	18	3			
$21 \le d < 26$	23	5			
$26 \le d < 31$	28	4			
$31 \le d < 36$	33	3			
$36 \le d < 41$	38	2			
	Totals	$\sum_{i=1}^{7} f_i = 20$		$\sum_{i=1}^{7} f_i x_i =$	$\sum_{i=1}^{7} f_i x_i^2$

Example 6: A restaurant serves a variety of wines of different prices. In a week chosen at random the number of bottles of wine sold was recorded by price and the results are shown in the table.

Price in euro, x	Number of bottles of wine sold				
$0 \le x < 5$	27				
$5 \le x < 10$	5				
$10 \le x < 15$	8				
$15 \le x < 20$	3				
$20 \le x < 25$	1				

Calculate the mean and the standard deviation. Verify your calculations using the GDC.

GDC INSTRUCTIONS FOR 1-VARIABLE STATISTICS <u>ON GROUPED DATA</u>:



- Press [STAT]
- Select [1] (for EDIT)
- Enter the **mid-value of the class interval in L1** and the **frequency in L2**.
- Press [STAT]
- Select the CALC men
- Select [1] (for 1-VAR Stats)
- Press [2nd] [1] [,] [2nd] [2] followed by [ENTER].

Empirical Rule

The empirical rule applies only to a specific type of distribution called a bell-shaped distribution, as shown in Figure 1. Such a distribution is called a normal curve.
In this section, only the following three rules for the curve are given.

Empirical Rule For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.7% of the observations lie within three standard deviations of the mean.

The empirical rule applies to both population data and sample data.

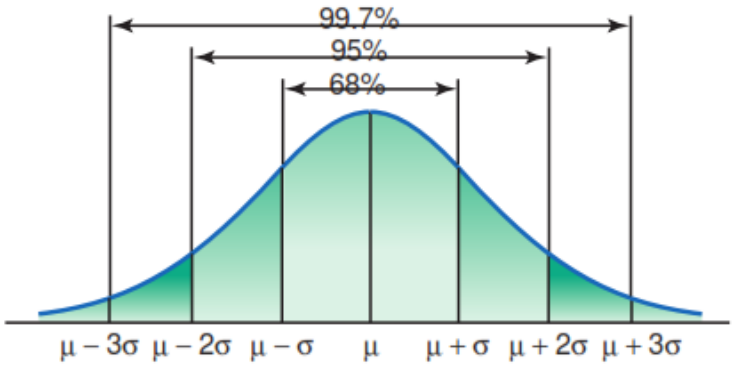


Figure 1.

Note: The **points of inflection** of the normal distribution curve occur at $\mu - \sigma$ and $\mu + \sigma$

Example 7: The age distribution of a sample of 5000 persons is bell shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.

Effect Of Constant Changes To The Original Data

If you add/subtract a constant value k to or from all the numbers in a list, the arithmetic mean increases/decreases by k but the standard deviation remains the same.

If you multiply or divide all the numbers in the list by a constant value of k, both the arithmetic mean and the standard deviation are multiplied or divided by k, respectively.

Multiplying all the numbers in the list by a constant value k, the variance will increase by the square of the constant.

For example: Given the values of a data set: 30, 17, 32, 25, 31, 28, 35, 26.

$$\bar{x} = \frac{\sum x}{n} = \frac{30+17+32+25+31+28+35+26}{8}$$
$$= 28$$

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n} = \frac{(30-28)^2+(17-28)^2+\cdots+(26-28)^2}{8}$$
$$\approx 26.5$$

$$\sigma \approx 5.15$$

Adding a constant of 2 to the list of numbers will result in: 32, 19, 34, 27, 33, 30, 37, 28.

$$\bar{x} = \frac{\sum x}{n} = \frac{(30+17+32+25+31+28+35+26)+8(2)}{8}$$
$$= 30$$

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n} = \frac{(32-30)^2+(19-30)^2+\cdots+(28-30)^2}{8}$$
$$\approx 26.5$$

$$\sigma \approx 5.15$$

Multiplying the list of numbers by a factor of 2 will result in: 60, 34, 64, 50, 62, 56, 70, 52.

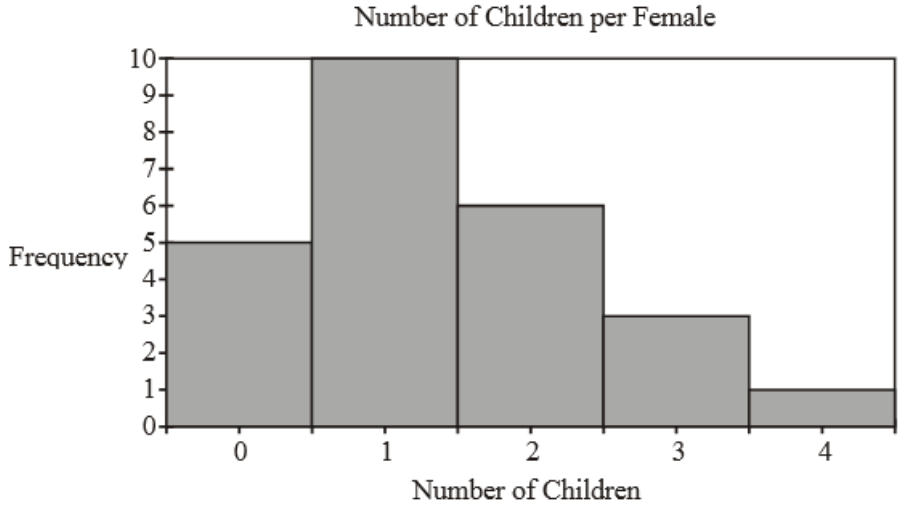
$$\bar{x} = \frac{\sum x}{n} = \frac{2(30+17+32+25+31+28+35+26)}{8}$$
$$= 56$$

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n} = \frac{(60-56)^2+(34-56)^2+\cdots+(52-56)^2}{8}$$
$$= \frac{[2(30-28)]^2+[2(17-28)]^2+\cdots+[2(26-28)]^2}{8}$$
$$= \frac{2^2\left[(30-28)^2+(17-28)^2+\cdots+(26-28)^2\right]}{8}$$
$$\approx 4(26.5)$$
$$\approx 106$$

$$\sigma \approx 10.3$$

Exit Card!

A group of 25 females were asked how many children they each had. The results are shown in the histogram below.



(a) Show that the mean number of children per female is 1.4.

(b) Show clearly that the standard deviation for this data is approximately 1.06.

(c) Another group of 25 females was surveyed and it was found that the mean number of children per female was 2.4 and the standard deviation was 2. Use the results from parts (a) and (b) to describe the differences between the number of children the two groups of females have.

Practice

1. The SAT scores of 100 students have a mean of 975 and a standard deviation of 105. The GPAs of the same 100 students have a mean of 3.16 and a standard deviation of .22. Is the relative variation in SAT scores larger or smaller than that in GPAs?

2. Consider the following two data sets.
Data Set I: 12 25 37 8 41
Data Set II: 19 32 44 15 48
Note that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the standard deviation for each of these two data sets using the formula for sample data. Comment on the relationship between the two standard deviations.

3. Consider the following two data sets.
Data Set I: 4 8 15 9 11
Data Set II: 8 16 30 18 22
Note that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 2. Calculate the standard deviation for each of these two data sets using the formula for population data. Comment on the relationship between the two standard deviations.

4. A sample of 3000 observations has a mean of 82 and a standard deviation of 16. Using the empirical rule, find what percentage of the observations fall in the intervals $\mu \pm 1\sigma$, $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$.

5. A large population has a mean of 310 and a standard deviation of 37. Using the empirical rule, find what percentage of the observations fall in the intervals $\mu \pm 1\sigma$, $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$.

6. The prices of all college textbooks follow a bell-shaped distribution with a mean of $105 and a standard deviation of $20.

a) Using the empirical rule, find the percentage of all college textbooks with their prices between
i. $85 and $125 ii. $65 and $145

b) Using the empirical rule, find the interval that contains the prices of 99.7% of college textbooks.

Answer

1. Simply divide the standard deviation by the mean in each case to find the coefficient of variation.

$\frac{105}{975} \times 100 = 10.77\%\,(\text{SATs})$ $\frac{0.22}{3.16} \times 100 = 6.96\%\,(\text{GPAs})$

$\therefore$ **the relative variation in SAT scores** is **larger than that in GPAs**

2. Adding a constant to each value in a data set does not change the distance between values so the standard deviation remains the same.

3. Multiplying a random variable by a constant increases the variance by the square of the constant.

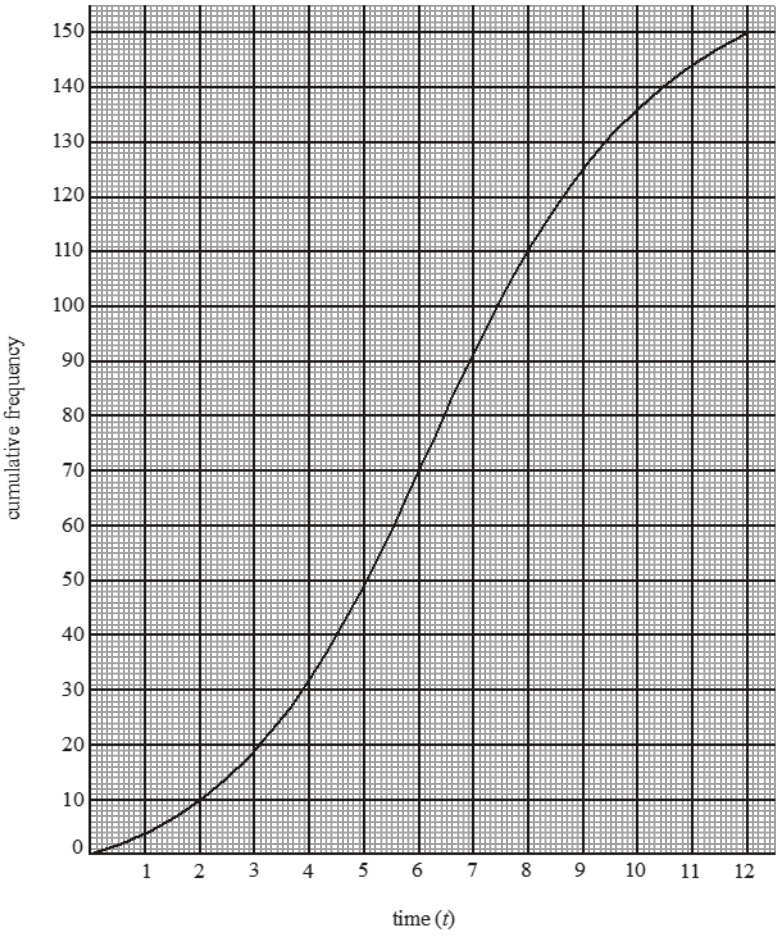
6. a) i. $\mu - k\sigma = 125 \rightarrow 105 - k(20) = 125 \rightarrow k = 1$ 68% will be in the interval $85 and $125

ii. $\mu - k\sigma = 145 \rightarrow 105 - k(20) = 145 \rightarrow k = 2$ 95% will be in the interval $65 and $145

b) the price of 99.7% of the of college textbooks.will be in the interval $\mu \pm 3\sigma$.That is $45 and $165

Warm up

1. The following is the cumulative frequency curve for the time, t minutes, spent by 150 people in a store on a particular day.



(a) (i) How many people spent less than 5 minutes in the store?

(ii) Find the number of people who spent between 5 and 7 minutes in the store.

(iii) Find the median time spent in the store.

(b) Given that 40% of the people spent longer than k minutes, find the value of k.

(c) (i) Complete the following frequency table.

t (minutes)	$0 \le t < 2$	$2 \le t < 4$	$4 \le t < 6$	$6 \le t < 8$	$8 \le t < 10$	$10 \le t < 12$
Frequency	10	23				15

(ii) Hence, calculate an estimate for the mean time spent in the store.

2. The mean of the population $x_1, x_2, \ldots\ldots, x_{25}$ is m. Given that $\sum_{i=1}^{25} x_i = 300$ and

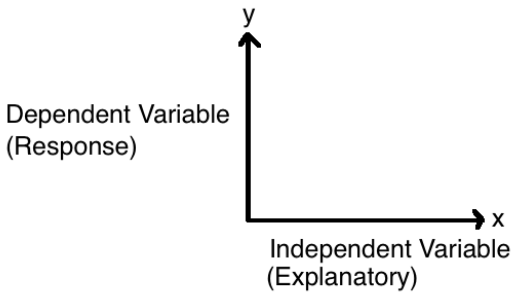
$\sum_{i=1}^{25} (x_i - m)^2 = 625$, find

(a) the value of m;

(b) the standard deviation of the population.

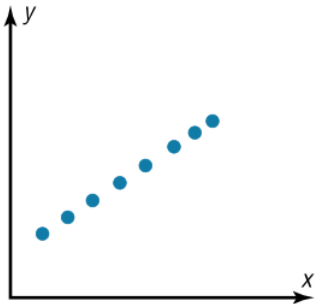
Linear Correlation and Regression

In data analysis, the relationship between variables is important. Statisticians often try to determine whether the dependent variable is affected by the independent variable.

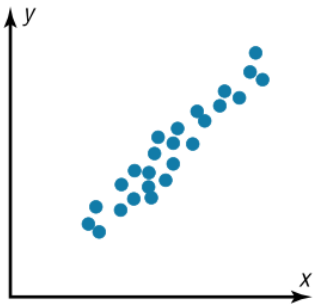


A **scatter plot** is a graph that describes the relationship between two variables. The **line of best fit** is the straight line that passes as close as possible to all the points on a scatter plot and represents the relationship between two variables.

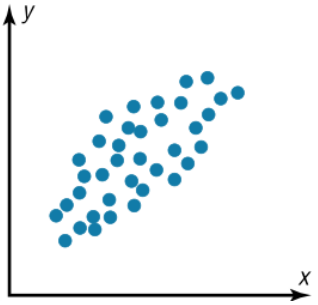
Variables have a **linear correlation** if changes in one variable are proportional to changes in another. The correlation may be described as positive or negative (direction of the relationship) and weak, moderate, or strong (strength of the relationship). It is also possible that the two variables have no correlation.



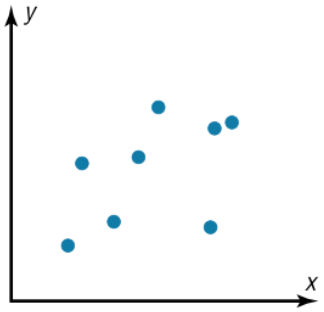
(a) Perfect positive Linear correlation



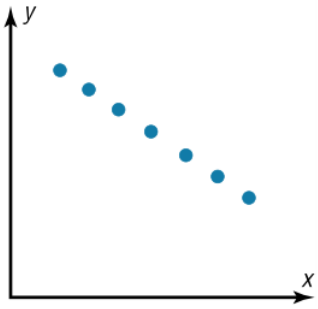
(b) Strong positive Linear correlation



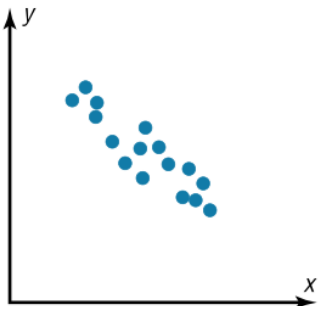
(c) Moderate positive Linear correlation



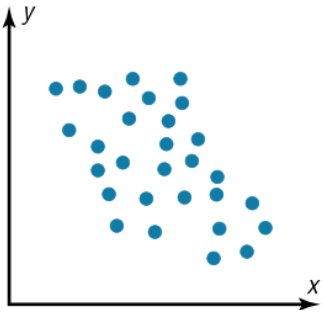
(d) Weak positive Linear correlation



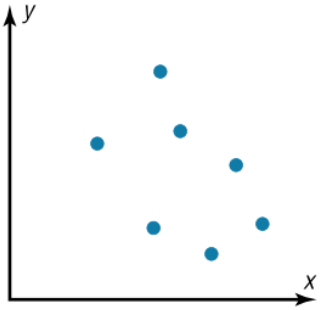
(a) Perfect negative Linear correlation



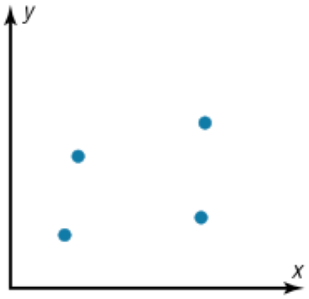
(b) Strong negative Linear correlation



(c) Moderate negative Linear correlation



(d) Weak negative Linear correlation

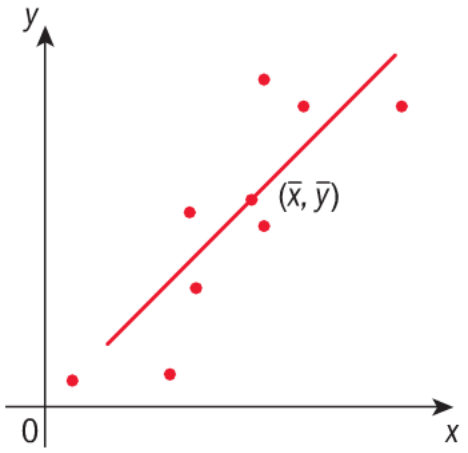


No Correlation

A correlation between two data sets does not necessarily mean that one causes the other. For example, there may be a strong positive correlation between ice cream consumption and number of drowning deaths in a given period of time. While these two variables may have nothing to do with each other, they appear strongly correlated. The reason is that there is a **confounding factor**, the season. The warmer weather in the summer leads to an increase in ice cream consumption as well as more people swimming and thus more drowning deaths.

Line of Best Fit and Mean Point

A **line of best fit** or **regression line** is drawn on a scatter diagram to find the direction of an association between two variables and to show the trend. This line of best fit can then be used to make predictions. However, it is likely that each person's line of best fit will differ from others. An improvement is to have a reference point for the line to pass through. This reference point is called the **mean point** $(\bar{x}, \bar{y})$ and is calculated by finding the mean of the x-values and the mean of the y-values.



The equation of the line of best fit can be used for prediction purposes. One method to calculate the equation of the regression line is to use the mean point $(\bar{x}, \bar{y})$ and another data point (x_i, y_i) that the line of best fit passes through. The equation will be in the form $\hat{y} = ax + b$, where a is the slope ($a = \frac{\bar{y} - y_i}{\bar{x} - x_i}$) and b is the y-intercept ($b = \bar{y} - a\bar{x}$).

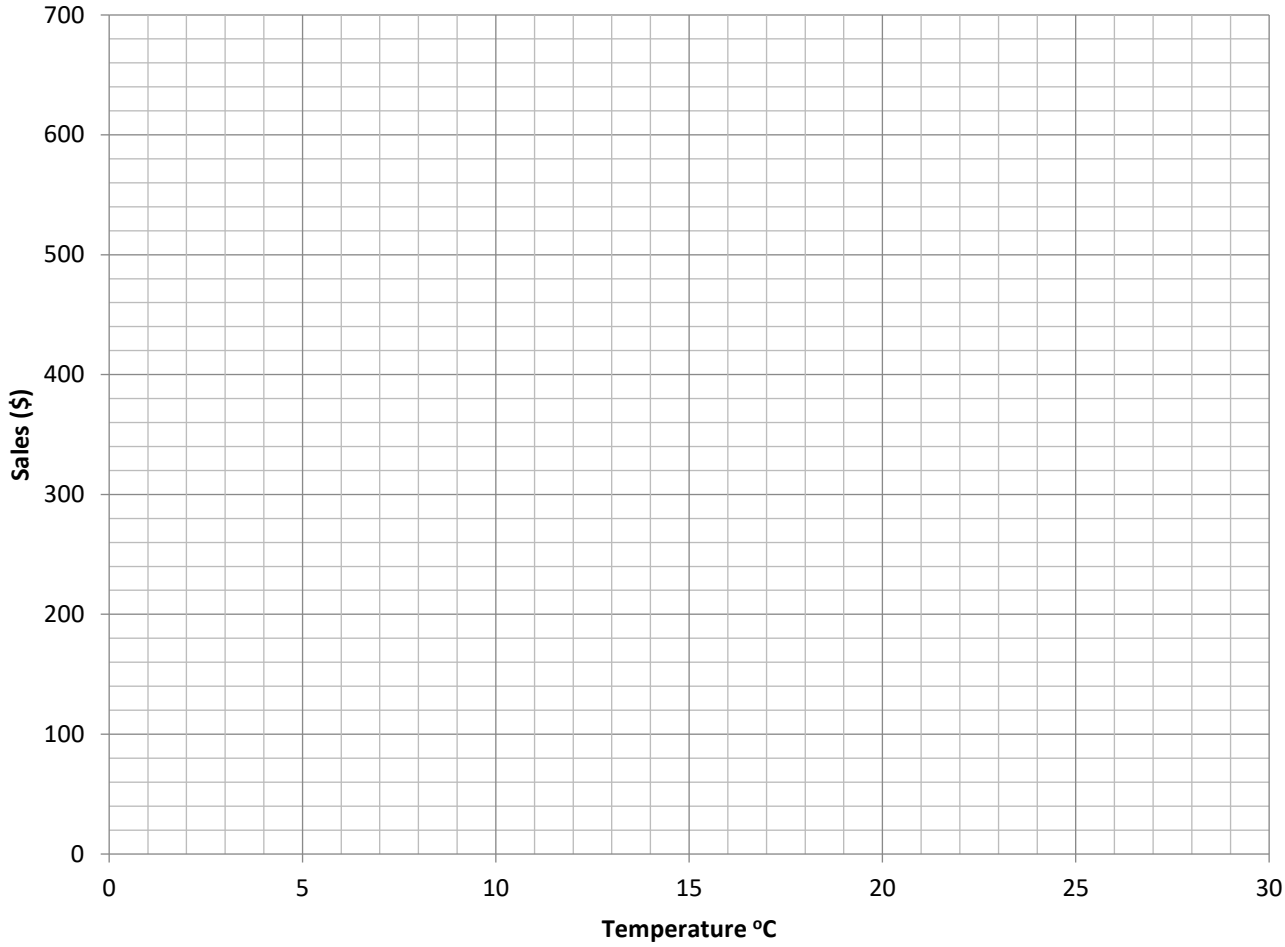
Example 1: The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day. Here are their figures for the last 8 days:

Temperature (°C)	14.2	16.4	11.9	15.2	18.5	22.1	19.4	25.1
Sales ($)	215	325	185	332	406	522	412	614

a) Find the mean temperature.

b) Find the mean amount of sales.

c) Construct a scatter plot for this data.



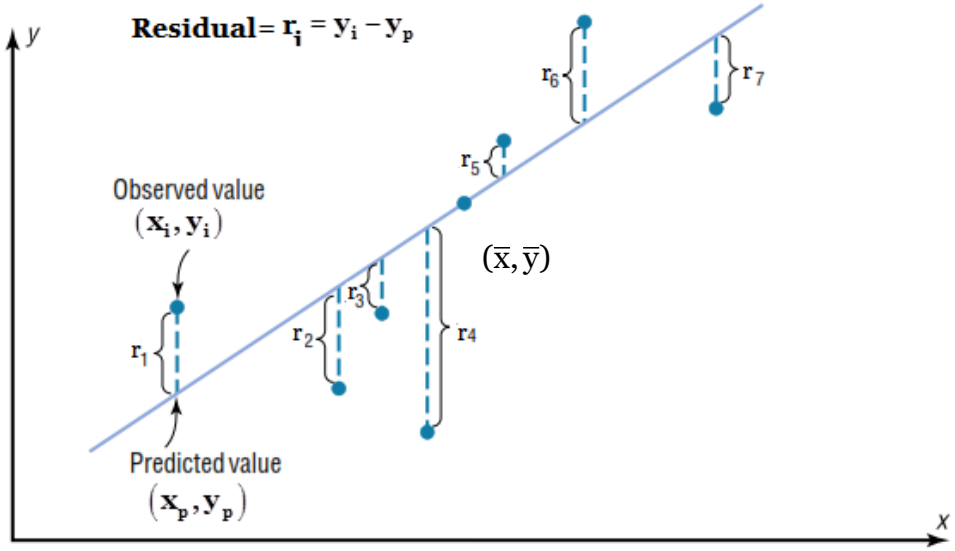
d) Plot the mean point on your scatter plot and use it to draw a line of best fit.

e) Determine the equation of the regression line.

Least Squares[1] Regression

Another objective way of producing the equation of a regression line is called the **least squares method**. The least squares method produces a straight line in which the sum of the squares that each point is from the line of best fit is minimized. The line always passes through the mean point, $(\bar{x}, \bar{y})$.

The **residual** of each data is the vertical distance between the data point and the line of best fit. The residual is positive if the data point is above the graph. The residual is negative if the data point is below the graph. The residual is zero when the graph passes through the data point.

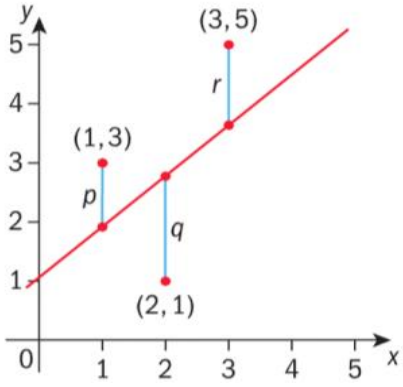


For the line of best fit using the least squares method:

- the sum of the residuals is zero
- the sum of the squares of the residuals has the least possible value

Example 1

In the diagram for example 1 to the right, $p^2 + q^2 + r^2$ should be as close to zero as possible.



Formulas for the Regression Line $\hat{y} = ax + b$,

$$a = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \quad , \quad b = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

[1]Francis Galton drew the line of best fit visually. An assistant of Karl Pearson's named G. Yule devised the mathematical solution using the least-squares method, employing a mathematical technique developed by Adrien-Marie Legendre about 100 years earlier.

Example 2. Use the least squares regression formula to determine the equation of the regression line through the data points in example 1 (above).

x	y	x^2	xy
1	3		
2	1		
3	5		
$\sum x$	$\sum y$	$\sum x^2$	$\sum xy$

Interpolation
Interpolation is when the model (the equation of the linear regression) is used to predict outcomes that are not included in the data set, but whose values are between the minimum and maximum values of the data set studied.

Extrapolation
Extrapolation is when the model (the equation of the linear regression) is used to predict outcomes for a data value that is not between the minimum and maximum of the data set.

USING GDC TO DETERMINE THE EQUATION OF THE REGRESSION LINE

- *Entering Data*
  - Press STAT 1 or STAT ENTER.
  - To place the data in L_1, enter the data and press ENTER to move to the next row. Press the right cursor to move to the L_2 column and continue to enter all the values.
- *Creating a Scatter Plot*
  - Press 2nd Y= 1 or 2nd Y= ENTER to select "Plot 1".
  - Highlight "Plot 1", "On" and "Type" (scatter plot). To do this, move the cursor on the selection and press ENTER.
  - Press ZOOM 9 to display the scatter plot.
- Turn Diagnostic ON (for the coefficient of correlation)
  - Press 2nd 0 for the catalog menu
  - Scroll down to DiagnosticON and press ENTER
- *Creating a Line of Best Fit*
  - Press STAT and cursor to "CALC".
  - Press 4 to select "LinReg(ax+b)". **DO NOT PRESS ENTER.**
  - Press 2nd 1 , 2nd 2 , VARS and cursor to "Y-VARS" and press 1 to select "Function", then press 1 to select "Y_1".
  - Press ENTER. The values on the screen make up the equation of the line of best fit and are stored in Y_1. This screen also displays the r and r^2 value. The r value represents the coefficient of correlation.
  - Press GRAPH to view the scatter plot with its corresponding line of best fit.

Example 3.

a) Use the Least Squares Method to find the equation of a line of best fit using GDC

Age (years) - (x)	Annual Income (1000$) - (Y)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35

b) Predict the income for an employee who is 21 and an employee retiring at age 65.

i) For a 21 year old employee, ii) For a 65 year old employee,

c) Determine which one is interpolation and which one is extrapolation from question b).

d) Comment on the accuracy of both estimates.

Question 1) The best-fitting line to a data set is given by $y=5.25x-2.84$. If the mean for y is $\bar{y}=8.75$, what is the mean for x *?*

a) $\bar{x}=2.21$
b) $\bar{x}=-2.21$
c) $\bar{x}=1.13$
d) $\bar{x}=43.1$

Question 2) The scores of ten students are collected from their English and their history exams, and are shown below.

English	54	59	50	75	82	67	71	73	44	60
History	65	71	63	78	87	66	70	76	58	65

A student sits the English exam and scores 62%. Unfortunately, the student was absent on the day of the history exam. Use the line of best fit to predict what that student might have scored on the history exam, to the nearest per cent.

a) 59%
b) 62%
c) 65%
d) 69%

Question 3) The equation of a line of best fit can be used to predict the value of a second variable when:

a)
- The correlation coefficient is close to positive one or close to negative one.
- The value being used is between the minimum and maximum data value studied.

b) The precise value already occurs in the data set.

c) The scatter diagram looks more like a cloud.

d)
- The second variable is caused by the first variable.
- The second variable is less than 100.

Question 4) Given the bivariate data in the table below, what is the value of y for the mean of x, i.e. $\bar{x}$, and does this require interpolation or extrapolation?

a) y=4.67, interpolation
b) y=5.04, interpolation
c) y=4.1, interpolation
d) y=−4.67, extrapolation

x	y
1	9
2	12
3	7
4	6
5	3
6	4
7	3
8	4
9	2
10	1
11	2
12	3

Measuring Linear Correlation

When considering a scatter plot, the direction and strength of the relationship between the two variables is of particular interest. The **coefficient of correlation, r,** is a numerical measure that describes the strength between the two variables. It is also known as the Pearson product-moment correlation coefficient (named after Karl Pearson) and has a value between -1 and +1 inclusive. The value reveals the strength of the correlation between the two variables.

Pearson's *r* correlation coefficient:

Ranges of Pearson product-moment correlation.	
Value of ***r***	**Description**
$0.7<r\leqslant 1$	Strong positive correlation
$0.3<r\leqslant 0.7$	Weak to moderate positive correlation
$-0.3<r\leqslant 0.3$	No correlation
$-0.7<r\leqslant -0.3$	Weak to moderate negative correlation
$-1\leqslant r\leqslant -0.7$	Strong negative correlation

Note : These intervals are very subjective and will vary depending on the textbook or statistician.

Linear Correlation Coefficient Formula:

The Pearson product-moment correlation coefficient can be found by hand, but it can be a very tedious process and, like standard deviation, is a prime example of the benefits of using technology.

The value of *r* is calculated using the following formula:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}}$$

An alternative version is the formula:

$$r = \frac{\sum xy - n\,\bar{x}\,\bar{y}}{\sqrt{\left[\sum x^2 - n\bar{x}^2\right]}\sqrt{\left[\sum y^2 - n\bar{y}^2\right]}}$$

or

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

Example 1. Determine the equation of the linear regression and the coefficient of correlation for the following:

a)

x	y	x^2	y^2	xy
2	13			
6	20			
7	27			
$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$

b)

x	y	x^2	y^2	xy
2	27			
6	20			
7	13			
$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$

Example 2:
A group of students are asked to jump from standing, first horizontally and then vertically. The vertical measurement is the distance between their reach and their jump. The following measurements were obtained:

Horizontal jump (cm)	Vertical jump (cm)
215	46
200	43
168	46
191	57
240	49
128	23
112	25
121	28
150	34
140	30
261	62
170	37
232	58
212	52
162	23

a) Determine a linear regression model equation using GDC to represent this data.

b) Decide whether the regression model is a "good fit" to represent this data. Justify your answer.

c) Comment on the meaning of the slope and y-intercept of $\hat{y}$.

Coefficient of Determination, R^2

R^2 measures the proportion of the variation in y that can be attributed/explained by the regression model generated (linear, quadratic, exponential, logarithmic, sinusoidal, etc.)

$$R^2 = \frac{explained\ variation}{total\ variation} = \frac{\sum\left(Y_{est} - Y_{avg}\right)^2}{\sum\left(Y - Y_{avg}\right)^2} or \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}$$

For linear regression, it turns out that $R^2 = r^2$

Note: The coefficient of determination, R^2, is used for <u>non-linear regression</u>

- $0 \leq R^2 \leq 1$, if $R^2 = 1$ the curve is a perfect fit of the data. If R^2 is greater than 0.8, the model has a good fit (this does not mean the model is good) and can be used to calculate reliable predictions of the dependent variable by using the independent variable.
- R^2 does not indicate whether the correct regression was used! R^2can be high for both very bad and very good models. End behaviours must be considered!
- R^2 does not indicate whether the independent variables are a cause of the changes in the dependent variable nor whether the most appropriate set of independent variables were chosen to measure the dependent variable.

Question 1) Match the scatter plots, Figures 1–4, with the following correlation coefficients: A: r=−0.81, B: r=−0.50, C: r=0.55 and D: r=0.96.

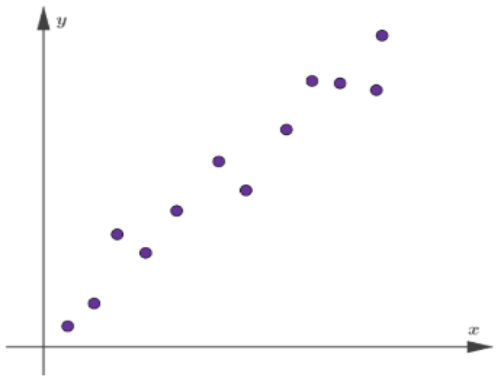


Figure 1.

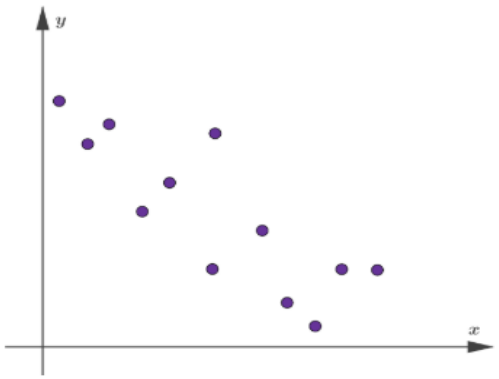


Figure 2.

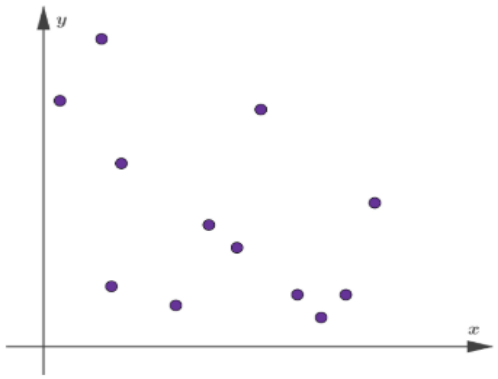


Figure 3.

Figure 4.

a) A: Figure 2, B: Figure 3, C: Figure 4, D: Figure 1
b) A: Figure 3, B: Figure 2, C: Figure 4, D: Figure 1
c) A: Figure 2, B: Figure 3, C: Figure 1, D: Figure 4
d) A: Figure 2, B: Figure 4, C: Figure 3, D: Figure 1

Question 2) Alejandro is exploring the connection between the heights (x) in centimetres of his berry bushes and the kilograms of berries (y) he harvested from each. The data he collected is shown in the table below. Find the correlation coefficient for the data

x (cm)	y(kg)
90	1.1
100	2.8
110	4.3
120	5.9
130	8.1
140	9.9

a) *0.176*
b) *0.998*
c) *0.996*
d) *−14.9*

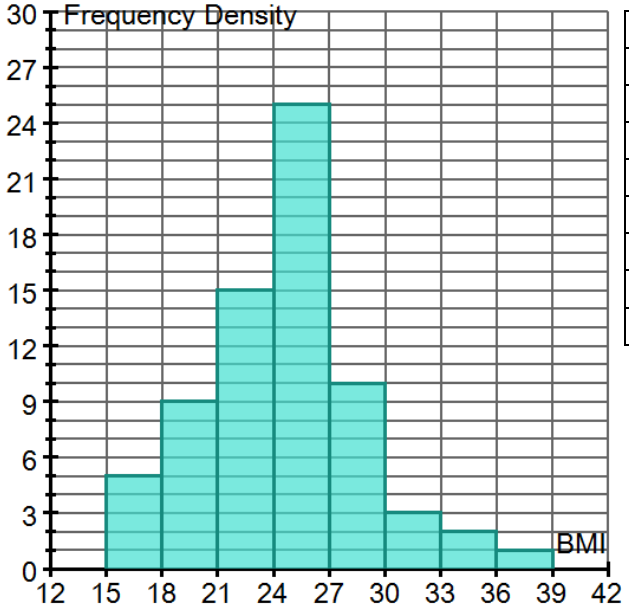
Descriptive Statistics Review Questions

1. Using the following stem and leaf plot, determine the mean, standard deviation, median, and mode given that $N = 34,\ \sum x = 810$, and $\sum x^2 = 25476$. $[\mu = 23.82, \sigma = 13.48,\ Q2 = 25.5,\ mode = 3]$

Stem	Leaf
0	0 3 3 3 4
1	0 0 1 1 2 3 5
2	3 3 4 5 5 6 7 7 8 9 9
3	0 5 5 6 8 9 9
4	3 4 4 6

2. Kevin works at LA Fitness and wants to assess the overall fitness of the gym's membership. He goes onto the gym's database and realizes that all of the member's information is confidential. He decides to look at some presentation files that the Gym's supervisor had prepared and notices a frequency density histogram used to display the membership data.

Using the Frequency Density histogram shown below he decides to perform some statistical analysis. Fill in the chart and answer the questions below. Note: The units of BMI are kg/m^2.



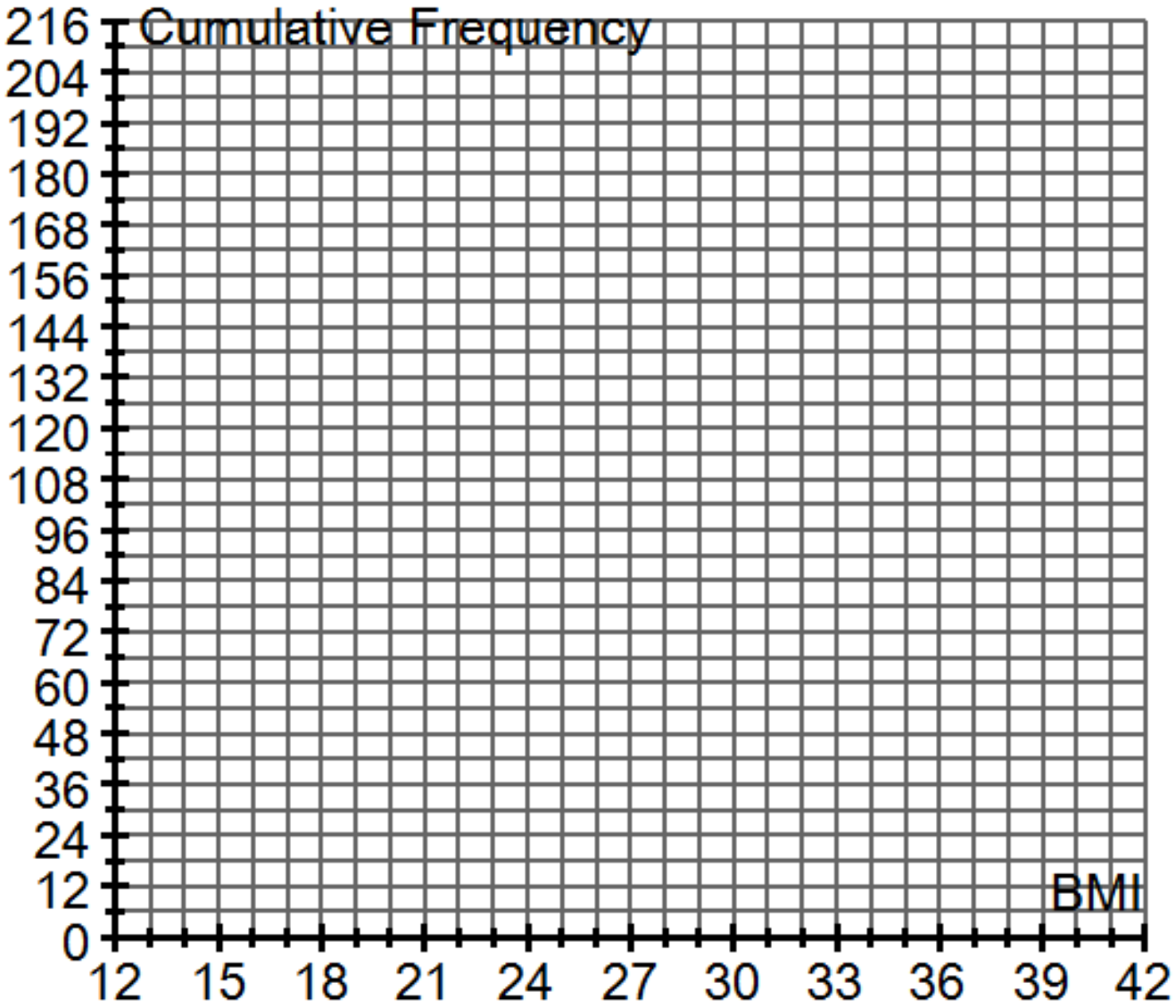
BMI Range	Frequency	Mid Values	Cumulative Frequency
$15 \le x < 18$			
$18 \le x < 21$			
$21 \le x < 24$			
$24 \le x < 27$			
$27 \le x < 30$			
$30 \le x < 33$			
$33 \le x < 36$			
$36 \le x < 39$			210

a. Provide an estimate to the mean and standard deviation of the gym's membership. $[\mu = 24.5571, \sigma = 4.319]$

b. A BMI is an attempt to quantify the amount of tissue mass (muscle, fat, and bone) in an individual and categorize them as being under weight, normal weight, over weight, and obese. The following table summarizes BMI. Using the table, what can you infer about the gym's membership?

BMI	Classification
$x < 18.5$	Under Weight
$18.5 \leq x < 25$	Normal Weight
$25 \leq x < 30$	Over Weight
$x \geq 30$	Obese

c. Using the data from your chart, construct a cumulative frequency curve on the grid provided.



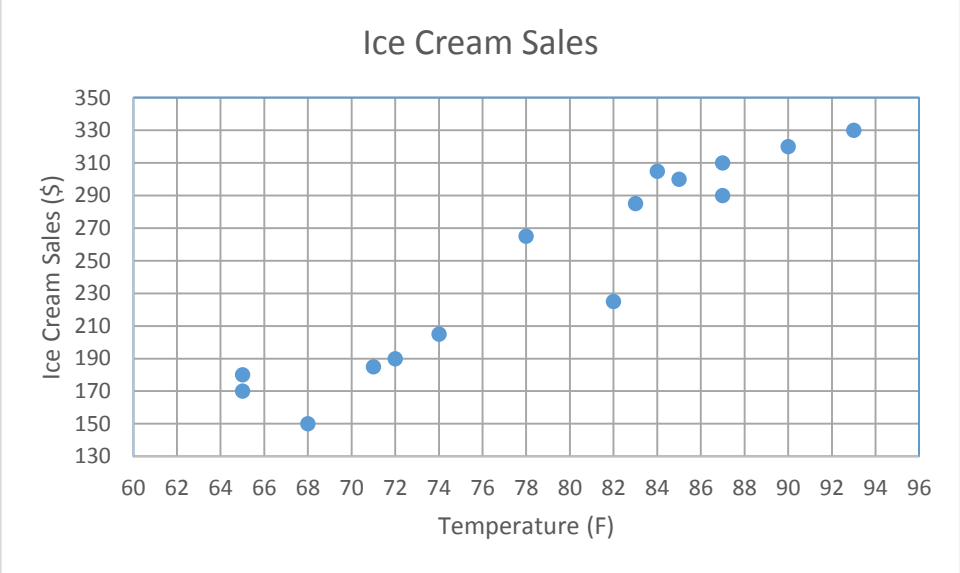
d. Using your cumulative frequency graph, estimate the locations of Q1, Q2, Q3 and use these values to construct a box and whisker plot. $[Q1 = 21.7, Q2 = 24.72, Q3 = 26.82 \ldots answers\ may\ varry]$

e. Using your calculations above, calculate the Lower and Upper fences and justify whether or not there may be an outlier in the dataset.

f. If there are no outliers in the dataset and the dataset is sufficiently large, what should you realize when looking at the mean and the median value?

3. An Ice Cream Salesman decided to analyze how his daily ice cream sales for the day were dependent on the temperature outside. Throughout the summer, he recorded his data. He then randomly selected 15 days to analyze his sales. Below are the results

Temperature	Ice Cream Sales
85	300
74	205
87	290
93	330
71	185
68	150
87	310
83	285
65	170
90	320
82	225
72	190
65	180
78	265
84	305



a) Using the GDC, calculate the linear regression model using least squares method.
$[\hat{y} = 6.4901x - 264.95,\ 65 \leq x \leq 93]$

b) Calculate the correlation coefficient and describe the type and strength of the correlation.

c) Comment on whether or not the linear regression model is a good fit. $[r = 0.9506]$

d) What is the meaning of the slope?

e) At what temperature does the Ice Cream Man begin to earn money according to the line of best fit? Note that $T_c = \left(T_f - 32\right) * \frac{5}{9}$. Calculate the temperature in Celsius and comment on why this estimate is not reliable. $\left[T_f = 40.82\ F,\ no\right]$

f) If the temperature was 70 outside, how much money would the ice cream man expect to make? Is this a reliable estimate? $[\$189.36, yes]$

1. Given the following table, where the lower class limit of the first class is 10 and the class width is 2 and answer the following questions below.

Table 12

Three-Year Rate of Return of Mutual Funds (as of 10/31/07)

13.50	13.16	10.53	14.74	13.20	12.24	12.61	19.11
14.47	12.29	13.92	16.16	12.07	10.99	15.07	10.06
14.14	12.77	19.74	12.76	13.34	11.32	15.41	17.37
13.51	15.44	15.10	17.13	12.37	16.34	11.34	10.57
15.70	13.28	23.76	22.68	14.81	23.54	19.65	14.07

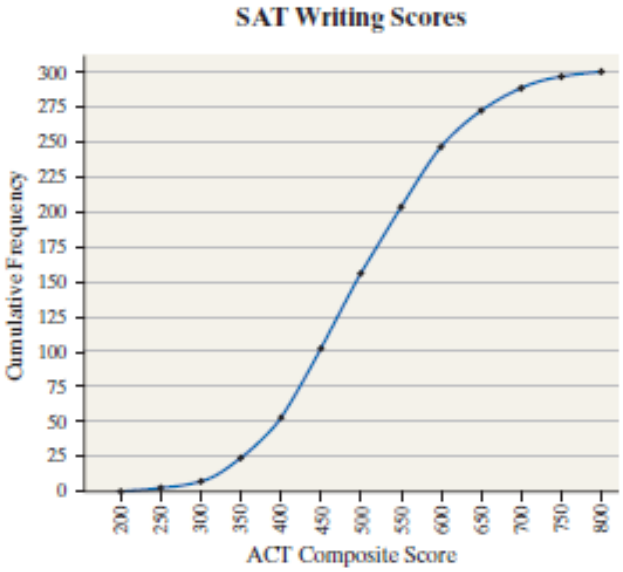
Source: TD Ameritrade

Class (3 year rate of return)	Frequency
$10 \leq x < 12$	6
$12 \leq x < 14$	14
	10
$22 \leq x < 24$	3

a) Draw a histogram of the data.
b) Calculate the mean and standard deviation of the raw data.
c) Calculate the mean and standard deviation using your frequency chart. $[13.8, 3.19]$
d) Find the modal class. $[12 \leq x < 14]$
e) Construct a cumulative frequency polygon for the data. (Use UCB as x coordinate)

2. The following cumulative frequency graph represents SAT writing scores of 300 randomly selected college-bound students

Source: Based on data from The College Board, *2007 College-Bound Seniors Total Group Profile Report*, 2007



a) What is the class width? $[50]$

b) How many classes are there? $[13]$

c) What are the lower and upper limits of the last class? $[750 \leq x < 800]$

d) Estimate the number of students who had a writing score of 350 or below. $[25]$

e) In which class did the most students fall? Estimate the number of students in this class. $[501 - 550, 54]$

f) Estimate Q1, Q2, Q3 and the IQR $[425, 490, 550]$

g) Draw a box and whisker plot of the data.

h) Determine if there are any outliers. $[\max \textit{and } \textit{min}]$

i) Determine the 4th Decal. What does this tell you? $[120]$

3. Data is gathered on a cars in a motor pool regarding the number of miles driven in a given year, and maintenance costs for that year. Here is the sample data

Car Number	Miles Driven (x) in thousands	Repair Costs (y)
1	80	$1200
2	29	$150
3	53	$650
4	13	$200
5	45	$325

a) Plot the data on the grid provided and use the mean point $(\overline{x}, \overline{y})$ to draw and estimate the line of best fit. $[(44, 505)]$

b) Using the line that you drew, find the equation of the line of best fit.

c) Calculate the line of best fit using Regression Analysis Formulas. $[y \doteq 16.0x - 197]$

d) Calculate Pearson's Correlation Coefficient and the Coefficient of Determination. What do each of these values tell you? $[r = 0.930, r^2 = 0.865]$

e) What does the slope tell you?

f) Does the y-intercept make sense?

g) Estimate the repair costs if a person drives 75 thousand miles. Is this a valid estimate? Justify your answer. $[\$999.80, yes]$

h) Estimate the repair costs if a person drives 100 thousand miles. Is this a valid estimate? Justify your answer. $[\$1398.84, no]$